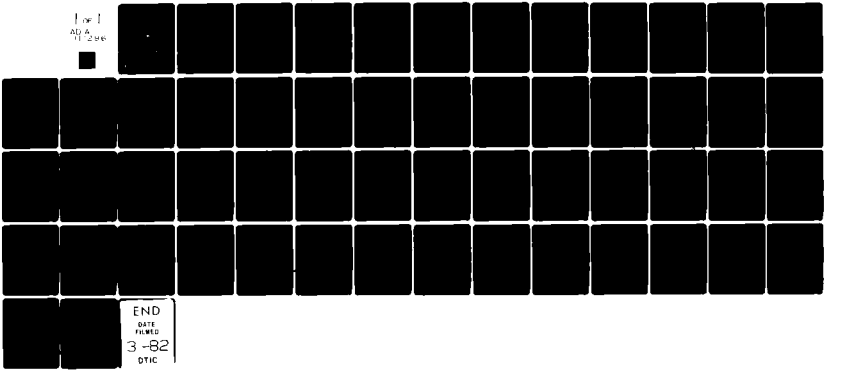
AD-A111 296 NAVAL BIODYNAMICS LAB NEW ORLEANS LA F/G 5/10
PERFORMANCE EVALUATION TESTS FOR ENVIRONMENTAL RESEARCH (PETER)--ETC(U)
JUL 81 R S KENNEDY, A C BITTNER, R C CARTER
UNCLASSIFIED NBDL-80R008 NL



30

NBDL - 80R008

PERFORMANCE EVALUATION TESTS FOR ENVIRONMENTAL RESEARCH (PETER): COLLECTED PAPERS

Robert S. Kennedy, Alvah C. Bittner, Jr., Robert C. Carter, Michele Krause, Mary M. Harbeson, Denise B. McCafferty, Ross L. Pepper, Steven F. Wiker

JULY 1981

NAVAL BIODYNAMICS LABORATORY
New Orleans, Louisiana

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER NBDL - 80R008	2. GOVT ACCESSION NO. AD-A111 276	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Performance Evaluation Tests for Environmental Research (PETER): Collected Papers		5. TYPE OF REPORT & PERIOD COVERED Research Report 6. PERFORMING ORG. REPORT NUMBER NBDL - 80R008
7. AUTHOR(s) R. Kennedy, A. Bittner, Jr., R. Carter, M. Krause, M. Harbeson, D. Mc Cafferty, R. Pepper, and S. Wiker		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS Naval Biodynamics Laboratory P.O. Box 29407 New Orleans, LA 70189		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS Project F58524 Task Area ZF5852406 Work Unit MF58.524-002-5027
11. CONTROLLING OFFICE NAME AND ADDRESS Naval Medical Research & Development Command Bethesda, MD 20014		12. REPORT DATE July 1981 13. NUMBER OF PAGES 43
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release, distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Repeated Measures	Memory	Digit span
Human Performance Testing	Item recognition	
Test Battery	Stroop	
PETER	Code substitution	

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This is a collection of papers about the ongoing development of Performance Evaluation Tests for Environmental Research (PETER). Environmental Research involves the assessment of human mental and physical capabilities in unusual environments (e.g., vibration, ship motion, deep sea diving, or outer space). Such research often includes repeated measurement of the capabilities of the same subjects before, during, and after exposure to an unusual environment. PETER is being developed specifically for repeated measurement, taking account of required properties of test means, variances, and intertrial correlations.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 68 IS OBSOLETE
S/N 0102-014-6601 |

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

(20 ABSTRACT)
Candidate tests for PETER were suggested by the literature of performance testing, as summarized in the first paper of this collection. The results of the examinations of candidate tests with respect to the required properties are summarized in the second paper of the collection. The remaining papers deal with specific tests, including code substitution, stroop, complex counting, critical tracking, time estimation, arithmetic, air combat maneuvering, digit span, four other memory tests, interference susceptibility, and item recognition, that are discussed only briefly in the first two papers. These tests were selected for no particular purpose, such as measuring specific attributes, rather they were selected based on availability and demonstrated usefulness. This collection of papers describes progress in the PETER project up to November 1980.

NBDL - 80R008

PERFORMANCE EVALUATION TESTS FOR ENVIRONMENTAL RESEARCH (PETER): COLLECTED PAPERS

Robert S. Kennedy, Alvah C. Bittner, Jr., Robert C. Carter, Michele Krause, Mary M. Harbeson, Denise B. McCafferty, Ross L. Pepper, Steven F. Wiker

July 1981

Bureau of Medicine and Surgery
Work Unit MF58.524-002-5027

Approved by

Channing L. Ewing, M. D.
Scientific Director

Released by

Captain J. E. Wenger MC USN
Commanding Officer

Naval Biodynamics Laboratory
Box 29407
New Orleans, LA 70189

Summary Page

PROBLEM

The effectiveness of many man-machine systems is limited by the performance of the human component. Environmental stressors, such as ship motion or vibration, are a major factor affecting human performance. Hence, it is important to know the degree to which performance capability is altered by environmental stressors encountered during operation of a man-machine system. Human performance capability can be assessed by comparing performance in a standard environment with performance in a stressful environment of interest. The comparison involves repeated measurement of the same subjects in both environments but not all performance tests are suitable for repeated measurement.

FINDINGS

1. Suitability of tests for repeated measurement can be represented by the means, variances, and intertrial correlations of test scores obtained from several measurements of the same subjects in a standard environment.

2. Tests become more suitable for repeated measurement after practice by the subjects. The required amount of practice varies from one test to another.

RECOMMENDATIONS

Tests that are to be used for repeated measurement should be practiced by the subjects prior to being used to obtain data. The required amount of practice should be determined from data obtained in a standard environment.

This research work was funded by the Naval Medical Research and Development Command and by the Biological Sciences Division of the Office of Naval Research.

The volunteers used in this study were recruited, evaluated and employed in accordance with the procedures specified in the Secretary of the Navy Instruction 3900.39 series and the Bureau of Medicine and Surgery Instruction 3900.6 series. These instructions are based upon voluntary consent, and meet or exceed the prevailing national and international guidelines.

Trade names of materials or products of commercial or non-government organizations are cited where essential for precision in describing research procedures or evaluation of results. Their use does not constitute official endorsement or approval of the use of such commercial hardware or software.

Table of Contents

Selection of Performance Evaluation Tests
for Environmental Research
by R. C. Carter, R. S. Kennedy, and A. C. Bittner, Jr. 1

A Catalogue of Performance Evaluation Tests
for Environmental Research
by R. S. Kennedy, R. C. Carter, and A. C. Bittner, Jr. 8

Performance Evaluation Tests for Environmental
Research (PETER): Code Substitution Test
by R. L. Pepper, R. S. Kennedy, A. C. Bittner, Jr.,
and S. F. Wiker 13

A Comparison of the Stroop Test to Other Tasks for
Studies of Environmental Stress
by M. M. Harbeson, R. S. Kennedy, and A. C. Bittner, Jr. 20

Performance Evaluation Tests for Environmental Research
(PETER): Auditory Digit Span
by D. B. McCafferty, A. C. Bittner, Jr., and R. C. Carter 29

Comparison of Memory Tests for Environmental Research
by M. M. Harbeson, M. Krause, and R. S. Kennedy 34

Performance Evaluation Tests for Environmental Research
(PETER): Interference Susceptibility Test (IST)
by M. Krause and R. S. Kennedy 41

Item Recognition as a Performance Evaluation Test for
Environmental Research
by R. C. Carter, R. S. Kennedy, A. C. Bittner, Jr.,
and M. Krause 47

Each of these papers was presented at a professional meeting or symposium. Acknowledgement of previous publication appears at the beginning of each paper.

PROCEEDINGS OF THE 24TH ANNUAL MEETING OF THE HUMAN FACTORS SOCIETY
LOS ANGELES, CA, 13-17 OCTOBER 1980

SELECTION OF PERFORMANCE EVALUATION TESTS FOR ENVIRONMENTAL RESEARCH[1]

Robert C. Carter, Robert S. Kennedy, and Alvah C. Bittner, Jr.
Naval Biodynamics Laboratory, New Orleans, LA 70189

ABSTRACT

A battery of Performance Evaluation Tests for Environmental Research (PETER) that is suitable for use in repeated measures experiments is being developed at the Naval Biodynamics Laboratory. This paper describes the sources of tasks which have been considered for inclusion in PETER. It also lists the tests in the source batteries which have or have not yet been considered for inclusion in PETER. The performance content of the tests that have been considered is compared with the content of those that have not. Recommendations are made for selection of additional tests from the source batteries which will not be redundant with tests that already have been considered. This report puts PETER into the context of the tests and test batteries which came before it.

INTRODUCTION

The Naval Biodynamics Laboratory is engaged in study of various measures of human performance in order to select Performance Evaluation Tests for Environmental Research (PETER) (Kennedy & Bittner, 1977; Kennedy, Bittner, & Harbeson, 1980). Several criteria have been used to choose the candidate tests. Prospective PETER tasks must have been shown to be diagnostic of brain damage, or to be sensitive to environmental stressors, or to measure some aspect of human information processing. Further, the test materials were required to be statistically suitable for repeated measurement of subjects' performance before, during, and after experiencing an unusual environment. In order to evaluate the suitability of tests, means, between-subject standard deviations, and cross-session reliabilities were obtained from 15 days of repeated measures in a standard environment. After a reasonable amount of practice, the means, standard deviations, and reliabilities must have been approximately constant across sessions. Constant means in a standard environment are preferred if changes due to an unusual environment are to be interpretable (Campbell & Stanley, 1966), although linearly-increasing means are also acceptable. Constant standard deviations and cross-session reliabilities are sufficient to meet some assumptions of repeated measures ANOVA (Winer, 1971), which is often employed to analyze environmental experiments. These, then, were the criteria for suitability of a test for assessment of performance in exotic environments. The purposes of this report are: (1) to show the sources of tests which have been considered for PETER; and (2) to evolve plans for the selection of additional tests.

METHOD

Candidate tests for PETER have been selected mainly from other performance test batteries because of the intellectual and financial investment in those batteries and the need for use of standardized procedures. The sources from which tests have been adopted for PETER include: Wechsler (1958); Ekstrom, French, Harman and Derman (1976)[2]; Fleishman and Ellison (1962); Rose (1974, 1978); Reitan and Davison (1974); Bennett (1979); Underwood, Boruch, and Malmi (1977); Video games, and other miscellaneous sources.

Many tasks within these batteries have not yet been considered for inclusion in PETER. In some cases, tasks were not adaptable for repeated measurement. For example, it would be almost impossible to generate many comparable forms of an information test (e.g. Wechsler, 1958). Numerous tests have not been examined because of necessary compromises involving resources available and judgements of the importance and uniqueness of test content.

Other batteries have not been studied for various reasons. For instance, Fleishman's (1964) tests of physical fitness have not been investigated because their scores are likely to change radically with repeated measurement. The extensive research of Alluisi (e.g. 1966) and others on synthetic work is not yet reflected in PETER because of the need to demonstrate suitability of component tasks before combining the tasks. In addition, batteries intended primarily for selection or training evaluation were not used because they are usually proprietary, and because they are more likely to measure success or achievement than performance. Finally, some performance batteries may have been unintentionally overlooked.

A tabular approach was employed to compare PETER with the source batteries and to aid selection of new tests for possible inclusion in PETER. One table was constructed for each source battery. The tables give the names of the tests in the battery and the performance functions measured by those tests. The tests listed in each table are classified as having been considered for inclusion in PETER or not. Hence, the tables fulfilled our first objective by showing the overlap between PETER and other test batteries. The second objective, selection of additional

[1] This research was performed under Navy Work Unit No. MF58.524.002-4027. The opinions are those of the authors and do not necessarily reflect those of the Department of the Navy.

[2] Tasks drawn from the Ekstrom et al. (1976) battery or its predecessors (e.g. Moran, Kimble, & Mefferd, 1964) are listed under this reference.

tests for PETER, was met by examining the tests which have not yet been considered for inclusion in PETER. Those tests which measure content not now represented in PETER were recommended for consideration.

RESULTS

Table 1 displays the tests in the Wechsler (1958) Adult Intelligence Scale, which is intended to measure ability to think rationally, to act purposefully, and to deal effectively with the environment. Three of the 11 tests have been entertained for inclusion in PETER (Arithmetic, Digit Span, and Code Substitution). It is obvious that they were chosen because alternate forms are relatively easy to generate. The remaining 8 tests, which have not yet been considered, measure range of experience (Information and Vocabulary), and ability to analyze and synthesize complex situations (Picture Completion, Comprehension, Similarities, Block Design, Picture Arrangement, and Object Assembly). It is apparent that we have reviewed the atomistic elements of the Wechsler battery, and have not examined the molar elements. Furthermore, we have considered the symbolic tests and not the verbal and pictorial tests.

Table 2 shows the tests in the Ekstrom, French, Harman, and Dermen (1976) battery, some of which have been offered in 20 alternate forms by Moran, Kimble, and Mefferd (1964). The purpose of this factor-analytic battery of 72 cognitive tests is to provide research workers with a 23-factor reference system for comparison of studies on mental abilities. Table 2 shows that 9 of the 23 factors are represented by tests that have been considered for inclusion in PETER. However, 14 factors have not been represented in PETER by tests from Ekstrom, French, Harman, and Dermen (1976). The factors which are not represented in PETER by these tests have to do with: identifying visual configurations in noise (Speed-of-Closure), 4 Fluency factors that relate to rapidity of producing non-repetitive but related responses (e.g. list things that are red), Reasoning (Inductive, Logical and General), Memory (Associative and Visual), Visualization of objects assembled by rotation of their parts, Flexibility (Figural and Use, e.g., list unusual uses for a given common object), and Verbal Comprehension (e.g. Vocabulary).

Table 3 lists tests of manual dexterity analyzed by Fleishman and Ellison (1962). They show that their battery of 21 tests can be represented by 5 meaningful factors: Wrist-Finger Speed, Finger Dexterity, Speed of Arm Movement, Manual Dexterity, and Aiming. Three of these 5 factors are represented by tests that have been considered for inclusion in PETER, although Wrist-Finger Speed and Aiming are both represented only by a tapping test. Better measures of each of these two factors are suggested by Fleishman and Ellison (1962). Factors which are not represented in PETER are Finger Dexterity, and Speed of Arm Movement.

Table 4 reviews tests suggested by Rose (1974, 1978) as representative of human information processing. All of these tests have been considered for inclusion in PETER because of their construct validity and because Rose (1974, 1978) has suggested how to produce alternate forms.

Table 5 recounts the tests of the Halstead-Reitan batteries described by Reitan and Davison (1974). Only 1 test from this battery has been considered for inclusion in PETER. The purpose of these tests, as applied by Halstead and Reitan, is to provide a basis from which inferences may be made regarding the organic integrity of the brain. Most of the tests have been shown to be sensitive to brain damage (Reitan & Davison, 1974). It seems unlikely, however, that some of these (e.g. the aphasia screening test) would be sensitive to the range of variability encountered in normal subjects. Other tests in the battery (e.g. Critical Flicker Frequency, and Lateral Dominance Examination) appear to have little relation to the work-related abilities at which PETER is aimed. However, the battery offers some unusual tests which may be related to abilities that are occasionally useful (e.g. Speech Sounds Perception, Rhythm, Finger Oscillation, Steadiness, Ballistic Arm Tapping, Orientation, Sandpaper Test, or Tactile Form Recognition).

Table 6 reveals the tests included in the Duke University Environmental Battery (Bennett, 1979). This battery is of special interest because its purpose is similar to that of PETER: detect and identify changes in human abilities caused by unusual environments. The battery described by Bennett (1979) reflects a special interest in hyperbaric environments. Most of the tests in the battery have been discussed in this paper in connection with other batteries, although it includes a unique test of intentional tremor which has not yet been considered for inclusion in PETER.

Table 7 recalls the battery of 24 memory tests which was factor analyzed by Underwood, Boruch, and Malmi (1977). The contents of this battery should be well represented in PETER because memory plays a central role in human performance. Underwood, et al. (1977) found 5 meaningful factors that described most of the variance in scores on their tests. The factors, which tended to be related to the type of memory task rather than the type of material being remembered, were: Paired Associates, Free Recall, Memory Span, Recognition, and Discrimination. Tests of two of these factors, Paired Associates and Discrimination, have not yet been considered for inclusion in PETER.

Table 8 acknowledges that microcomputer-based video games have been considered as performance tests for possible inclusion in PETER (e.g. Kennedy, Bittner, & Jones, 1980). This source of tests is so new that it is difficult to compare its content with that of traditional tests. However, we have found that the Air Combat Man-

euvering game produces scores that are highly correlated with scores from traditional compensatory tracking. Furthermore, a recent factor analysis of five video games (the first 5 in Table 8) indicates that they are spanned by two factors represented by Air Combat Maneuvering and Slalom games (Kennedy, Bittner, & Jones, 1980). Such games will continue to be selected for consideration for inclusion in PETER.

Finally, Table 9 assembles some miscellaneous performance tests which have been examined but are not from an established battery. The Navigation Plotting test was selected because it is a task which is vital in the Naval context which motivates the development of PETER. The Landolt C test of visual acuity is the only sensory function test that has been considered for inclusion in PETER. Time estimation, multiple choice reaction time, and tracking (performed singly and in dual modes) were tried because of their prominence in the armamentarium of performance measurement.

DISCUSSION

Where do we go from here? It is obvious that the tests that have been considered for PETER are overwhelmingly representative of mental ability (i.e. throughput) tests. We believe, however, that tests of input and output capabilities should be included in PETER to supplement the tests of mental mediation. Some tests of visual perception, should be considered such as contrast sensitivity, dynamic visual acuity, color discrimination, accommodation, visual field size, fusional reserve, visual illusions, vection sensitivity, pattern recognition, and visual search. Tests of auditory perception may also be worthwhile, such as audiometry, the Rhythm test (Table 5), impedance audiometry, Naval Aviator's Speech Discrimination Test, and rhyming-word-list tests of speech perception. Tests of contaneous information processing suggested by Table 5 may also be of interest. In addition to these tests of input functions, some output tests may be of interest. For example, the most rudimentary form of output is standing erect. Tests of standing steadiness, postural tremor, and intentional tremor (such as the Ball Bearing Test in Table 6 or the steadiness test in Table 5) are examples of tests of fundamental output functions. Other tests of output were suggested by Table 3 which dealt with Fleishman and Ellison's (1962) manual performance factors. Tests of Finger Dexterity and Speed of Arm Movement are needed. The latter factor was also suggested by Reitan and Davison (1974) as represented in Table 5 (Ballistic Arm Tapping test). Additional Independent tests of Aiming and Wrist-Finger Speed also would be prudent selections.

Paired Associates was found to be the most influential factor in the Underwood, et al. (1977) analysis of memory. This factor, which is also reported by Ekstrom et al. (1976) is not represented by tests already considered for PETER. Other memory factors (from Tables 7 and 2, respectively) which are not represented by PETER tests are: (a) Discrimination (e.g. given many pairs of words, one of which is underlined in each pair, underline the appropriate word when one of the pairs is presented again), and (b) Visual Memory (e.g. reproduce a map).

Review of Table 2 showed that there were several families of cognitive factors that had not yet been considered for inclusion in PETER: Speed of Closure, Fluency, Reasoning, Visualization, and Flexibility. These are important determinants of human performance, and tests representing them should be investigated for inclusion in PETER.

Tests of human information processing (Table 4) offered by Rose (1974, 1978) have been exhaustively studied for inclusion in PETER. No additional tests of this type need be selected unless a new and important information processing paradigm becomes available. However, many of the tests already considered are ideally suited for implementation in a computer controlled form which may vastly improve the tests compared with the paper and pencil forms offered by Rose (1974, 1978).

Video games should continue to be selected for possible inclusion in PETER because they are adaptive, challenging, and interesting to perform. Interest in the task is very important when repeated measurements are to be made, as is common in environmental research. Furthermore, the dynamic nature of video games enables them to tap aspects of mental capability that are unavailable to paper and pencil tests and seemingly well related to operational jobs.

The global measures of performance offered by Wechsler (1958) and listed in Table 1 are largely not amenable to repeated measurement due to the difficulty of creating good alternate forms. Some of the performance factors measured by these tests may be assessed by other means. Range of experience could be represented by biographical items, for example. Ability to analyze and synthesize complex situations may be assessed with complex exercises such as war games.

To summarize, the following additional types of tests should be selected for possible inclusion in PETER:

1. Visual Perception
2. Auditory Perception
3. Tactile Perception
4. Standing Steadiness & Tremor
5. Finger Dexterity
6. Speed of Arm Movement
7. Aiming
8. Wrist-Finger Speed
9. Paired-Associates Memory
10. Discrimination Memory
11. Visual Memory
12. Speed of Closure
13. Fluency (4 types)

14. Reasoning (3 types)
15. Visualization
16. Flexibility (2 types)
17. Video games
18. Complex games requiring Analysis and Synthesis

Tests of these content areas are available in the source batteries discussed in this report, but such tests have not yet been considered for inclusion in PETER. It is recommended that attention be given to tests of these content areas.

REFERENCES

Alluisi, E. A. Methodology in the use of synthetic tests to assess complex performance, Human Factors, 1966, 9, 375-384.

Bennett, P. B. Personal communication, June 6, 1979.

Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research. Chicago: Rand McNally, 1966.

Ekstrom, R. B., French, J. W., Harman, H. H., & Dermen, D. Manual for kit of factor-referenced cognitive tests. Princeton, New Jersey: Educational Testing Service, 1976.

Fleishman, E. A. The structure and measurement of physical fitness. Englewood Cliffs, New Jersey: Prentice-Hall, 1964.

Fleishman, E. A., & Ellison, G. D. A factor analysis of fine manipulative tests. Journal of Applied Psychology, 1962, 46, 96-105.

Kennedy, R. S., & Bittner, Jr., A. C. The development of a Navy Performance Evaluation Test for Environmental Research (PETER). In L. T. Pope & D. Meister (Eds.), Productivity Enhancement: Personnel Performance Assessment in Navy Systems. Symposium presented at the Naval Personnel Research and Development Center, San Diego, October 1977, 393-408. (NTIS No. AD A056047)

Kennedy, R. S., Bittner, Jr., A. C., & Harbeson, M. M. An engineering approach to the standardization of Performance Evaluation Tests for Environmental Research (PETER). Proceedings of the 11th Annual Conference of the Environmental Design and Research Association (EDRA), Charleston, SC, March 1980.

Kennedy, R. S., Bittner, Jr., A. C., & Jones, M. B. The utility of available television-computer games for assessing performance and other applications. Proceedings of the 51st Annual Scientific Meeting of the Aerospace Medical Association, 63-64, May 1980.

Moran, L. J., Kimble, J. P., & Mefferd, R. B. Repetitive psychometric measures: Equating alternate forms. Psychological Reports, 1964, 14, 335-338.

Reitan, R. M., & Davison, L. A. Clinical Neuropsychology: current status and applications. New York: Halstead Preass, 1974.

Robb, G. P., Bernardoni, L. C., & Jonson, R. W. Assessment of Individual Mental Ability. New York: Intext Educational Publishers, 1972.

Rose, A. M. Human information processing: An assessment and research battery, Technical Report No. 46. Ann Arbor, Michigan: University of Michigan, January, 1974.

Rose, A. M. An information processing approach to performance assessment, AIR 58500-11/78-FR. Washington, D.C.: American Institutes for Research, 1978.

Underwood, B. J., Boruch, R. F., & Malmi, R. A. The composition of episodic memory, Evanston, Illinois: Northwestern University, 1977 (NTIS No. AD-040-696).

Wechsler, D. Measurement and appraisal of adult intelligence. Baltimore: Williams & Wilkins, 1958.

Winer, B. J. Statistical principles in experimental design (2nd ed.). New York: McGraw-Hill, 1971.

TABLE 1: WECHSLER (1958) TESTS AND PETER

TESTS	CONTENT[a]
CONSIDERED FOR PETER	
ARITHMETIC	ARITHMETIC PROCESSES
DIGIT SPAN	RETENTIVENESS, AUDITORY-IMAGERY, ATTENTION, AND CONCENTRATION
CODE SUBSTITUTION	ROTE RECALL, VISUAL IMAGERY, SPEED AND ACCURACY IN LEARNING AND WRITING SYMBOLS
NOT YET CONSIDERED	
INFORMATION	RANGE OF INFORMATION, EXPERIENCE
VOCABULARY	RANGE OF IDEAS, CONCEPT FORMATION LANGUAGE DEVELOPMENT
PICTURE COMPLETION	VISUAL IMAGERY, PERCEPTION AND ALERTNESS, CONCENTRATION
COMPREHENSION	SOCIAL JUDGEMENT, REASONING, ORGANIZATION AND APPLICATION OF KNOWLEDGE
SIMILARITIES	VERBAL CONCEPTS, ABSTRACT THINKING,
BLOCK DESIGN	FORM PERCEPTION, ANALYSIS AND SYNTHESIS
PICTURE ARRANGEMENT	ABILITY TO COMPREHEND A WHOLE SITUATION
OBJECT ASSEMBLY	VISUAL PERCEPTION AND SYNTHESIS, RECOGNITION OF PATTERNS

[a] Robb, Bernardoni, and Johnson (1972)

TABLE 2: EKSTROM, FRENCH, HARMAN, AND DERMAN (1976) AND MORAN, KIMBLE, AND MEFFERD (1964) TESTS AND PETER

TESTS	CONTENT
CONSIDERED FOR PETER	
COPYING[1]	FLEXIBILITY OF CLOSURE
HIDDEN WORDS[a]	VERBAL CLOSURE
WORD BEGINNINGS	WORD FLUENCY
CALENDAR TEST	INTEGRATIVE PROCESSES
AUDITORY DIGIT SPAN	MEMORY SPAN
ARITHMETIC OPERATIONS[a]	NUMBER
ADDITION	FACILITY
FINDING A, NUMBER COMPARISON, NUMBER CROSS OUT[1]	PERCEPTUAL SPEED
LETTER ROTATION	SPATIAL ORIENTATION
LINE FOLLOWING[a]	SPATIAL SCANNING
NOT YET CONSIDERED[b]	
E.G. SNOWY PICTURES	SPEED OF CLOSURE
E.G. OPPOSITES	ASSOCIATIONAL FLUENCY
E.G. MAKING SENTENCES	EXPRESSIONAL FLUENCY
E.G. ORNAMENTATION OF SIMPLE FIGURES	FIGURAL FLUENCY
E.G. LIST THINGS THAT SHARE A GIVEN CHARACTERISTIC	IDEATIONAL FLUENCY
E.G. INDUCE THE RULE IN A GROUP OF LETTER SETS	INDUCTION
E.G. FIRST AND LAST NAMES	ASSOCIATIVE MEMORY
E.G. MAP MEMORY	VISUAL MEMORY
E.G. NECESSARY ARITHMETIC OPERATIONS	GENERAL REASONING
E.G. SYLLOGISMS	LOGICAL REASONING
E.G. VOCABULARY	VERBAL COMPREHENSION
E.G. SURFACE DEVELOPMENT	VISUALIZATION
E.G. PLANNING PATTERNS	FIGURAL FLEXIBILITY
E.G. DIFFERENT USES OF COMMON OBJECTS	FLEXIBILITY OF USE

[1] Test forms from Moran, Kimble, and Mefferd (1964)
[a] One test is listed as an example of each cognitive factor which has not yet been conidered for inclusion in PETER.

TABLE 3: FLEISHMAN AND ELLISON (1962) MANUAL DEXTERITY TESTS AND PETER

TESTS		CONTENT
CONSIDERED FOR PETER		
AIMING (TAPPING SMALL CIRCLES)		AIMING, WRIST-FINGER SPEED
MINNESOTA RATE OF MANIPULATION:	PLACING	MANUAL DEXTERITY
	TURNING	MANUAL DEXTERITY
NOT YET CONSIDERED		**CONTENT**
MEDIUM TAPPING		WRIST-FINGER SPEED
LARGE TAPPING		WRIST-FINGER SPEED
PURSUIT AIMING: I, II		AIMING
SQUARE MARKING		UNIQUE
TRACING		UNIQUE
STEADINESS		UNIQUE
DISCRIMINATION REACTION TIME (PRINTED)		WRIST-FINTER SPEED, MANUAL DEXTERITY
PRECISION STEADINESS		UNIQUE
TEN-TARGET AIMING: ERRORS, CORRECTS		SPEED OF ARM MOVEMENT
HAND PRECISION AIMING: ERRORS, CORRECTS		SPEED OF ARM MOVEMENT
PIN STICK		FINGER DEXTERITY
PURDUE PEGBOARD		FINGER DEXTERITY
O'CONNOR FINGER DEXTERITY		FINGER DEXTERITY

TABLE 4: ROSE (1974, 1978) INFORMATION PROCESSING TESTS AND PETER

TEST	CONTENT
CONSIDERED FOR PETER	
LETTER ROTATION	ROTATION
NEISSER SEARCH	DECISION TIME
STERNBERG ITEM RECOGNITION	MEMORY SCANNING
LETTER RECALL (DIGIT SPAN)	ROTE MEMORY
MENTAL ADDITION	TRANSFORMATION, STORAGE, RETRIEVAL
GRAMMATICAL REASONING	VERBAL ABILITY
SEMANTIC MEMORY	ACCESS LONG TERM MEMORY
GRAPHEMIC & PHONEMIC ANALYSIS	ACCESS LONG TERM MEMORY
POSNER LETTER CLASSIFICATION	STORAGE AND RETRIEVAL
LEXICAL DECISION MAKING	ACCESS LONG TERM MEMORY
FITTS TAPPING	INFORMATION PROCESSING RATE
CRITICAL TRACKING	CONTROL LOOP DELAY
SERIAL TASKS	RESPONSE COMPLETION

TABLE 5: REITAN AND DAVISON (1974) TESTS AND PETER

TEST	CONTENT
CONSIDERED FOR PETER	
TRAIL MAKING	RAPID TAPPING IN A SPECIFIC PATTERN
NOT YET CONSIDERED	CONTENT
CATEGORY TEST	VISUAL FIGURE IDENTIFICATION
TACTUAL PERFORMANCE TEST	TACTILE FIGURE RECOGNITION, AND ASSEMBLY
RHYTHM TEST	COMPARISON OF RHYTHMIC SEQUENCES
SPEECH SOUNDS PERCEPTION TEST	DISCRIMINATE WORDS FROM ALTERNATIVES
FINGER OSCILLATION TEST	SPEED OF FINGER TAPPING
CRITICAL FLICKER FREQUENCY	FUSION OF A FLASHING LIGHT
STEADINESS BATTERY	COORDINATION AND TREMOR
LATERAL DOMINANCE EXAMINATION	HAND, FOOT, AND EYE DOMINANCE
[illegible] RANGE ACHIEVEMENT TEST	READING, SPELLING, ARITHMETIC
MINNESOTA MULTIPHASIC PERSONALITY INVENTORY	PERSONALITY
APHASIA SCREENING TEST	VERBAL EXPRESSION
BALLISTIC ARM TAPPING	LARGE ARM MOVEMENTS
ORIENTATION TEST	RIGHT-LEFT RECOGNITION & IDENTIFICATION
DYNAMOMETER	GRIP STRENGTH
SANDPAPER TEST	EVALUATE TEXTURE
VISUAL SPACE ROTATION	DRAW "X" WITH ROTATED VISION OF HAND
TACTILE FORM RECOGNITION TEST	TACTILE FORM RECOGNITION
PICTURE VOCABULARY TEST	GIVE NAMES OF PICTURE OBJECTS

TABLE 6: DUKE UNIVERSITY ENVIRONMENTAL BATTERY TESTS[a] AND PETER

TEST	CONTENT
CONSIDERED FOR PETER	
ARITHMETIC	NUMBER FACILITY
STROOP, COLOR, AND CONTROL	RESPONSE COMPETITION
BADDELEY'S GRAMMATICAL REASONING	VERBAL ABILITY
DIGIT SPAN	MEMORY
NUMBER COMPARISON	PERCEPTUAL SPEED
NOT YET CONSIDERED	CONTENT
BALL BEARING TEST	INTENTIONAL TREMOR
PURDUE PEG BOARD	FINGER DEXTERITY
BENNETT HAND TOOL DEXTERITY TEST	MANUAL DEXTERITY

[a] BENNETT (1979)

TABLE 7: UNDERWOOD, BORGER, AND BAUMEL (1977) TESTS OF MEMORY AND PETER

TEST	CONTENT
CONSIDERED FOR PETER	
FREE RECALL-CONTROL	FREE RECALL
FREE RECALL-CONCRETE WORDS	FREE RECALL
FREE RECALL-ABSTRACT WORDS	FREE RECALL
LIST DIFFERENTIATION	FREE RECALL
RUNNING RECOGNITION	RECOGNITION
DIGIT SPAN	MEMORY SPAN
INTERFERENCE SUSCEPTIBILITY	UNIQUE
NOT CONSIDERED FOR PETER	CONTENT
E.G. PAIRED ASSOCIATES, SERIAL LEARNING	PAIRED ASSOCIATES
E.G. VERBAL DISCRIMINATION	DISCRIMINATION

TABLE 8: ATARI® GAMES AND PETER

TEST	CONTENT
CONSIDERED FOR PETER	
AIR COMBAT MANEUVERING (ACM)	COMPENSATORY TRACKING
SLALOM	UNKNOWN
BREAKOUT	SAME AS ACM
RACECAR	SAME AS SLALOM
SURROUND	ACM AND SLALOM
ICE RACE	UNKNOWN
PONG	UNKNOWN
BASKETBALL	UNKNOWN
ANTI-AIRCRAFT	UNKNOWN
FLAG CAPTURE	UNKNOWN

TABLE 9: MISCELLANEOUS TESTS AND PETER

TEST	CONTENT
CONSIDERED FOR PETER	
NAVIGATION PLOTTING	MANEUVERING BOARD SOLUTIONS
LANDOLT C	VISUAL ACUITY
TIME ESTIMATION	CONTINUITY OF ATTENTION
MULTIPLE CHOICE REACTION TIME	REACTION TIME
DUAL CRITICAL TRACKING	TIME SHARING
COMPENSATORY TRACKING	TRACKING

PROCEEDINGS OF THE 24TH ANNUAL MEETING OF THE HUMAN FACTORS SOCIETY
LOS ANGELES, CA, 13-17 OCTOBER 1980

A CATALOGUE OF PERFORMANCE EVALUATION TESTS FOR ENVIRONMENTAL RESEARCH

Robert S. Kennedy, Robert C. Carter, and Alvah C. Bittner, Jr.
Naval Biodynamics Laboratory, New Orleans, LA 70189

ABSTRACT

Performance Evaluation Tests for Environmental Research (PETER) are under development at the Naval Biodynamics Laboratory and supporting organizations. The tests, or tasks, studied in this program have been largely derived from the literature. Each task was evaluated for suitability for repeated measures experimental designs which are almost universally used in environmental research. Suitability criteria included the "stability" of task means, standard deviations, and between trial correlations. The magnitude of the "stabilized" between-trial correlations, task definition, was also examined with respect to the administration time. There are 60 active tasks in the present program. All tasks examined to date exhibit stable means and variances after adequate practice but: (a) less than 30% meet minimal stability criteria for intertrial correlations; and (b) substantial practice (typically more than an hour over five days) is required to achieve stability. A tabular catalogue of the research findings and background for 15 tasks is presented and discussed.

INTRODUCTION

Background

An engineering approach to the development and standardization of a battery of Performance Evaluation Tests for Environmental Research (PETER) is underway under the direction of the Naval Biodynamics Laboratory. This approach involves test and evaluation of performance tasks prior to their being employed in the assessment of environmental effects. The goal of this effort is to ensure that selected tasks will be suitable for simple analysis and interpretation when employed in repeated-measures experiments (Kennedy & Bittner, 1977; Kennedy, Bittner & Harbeson, 1980). The emphasis is on statistical requirements for repeated-measures experimental designs because environmental research usually includes measurement of performance before, during, and after exposure to an unusual environment.

The criteria for suitable stability of tests used in repeated measures experiments have been delineated by Jones (1980) and Kennedy et al. (1980). These authors have suggested that "stability" exists when: (a) group mean performance in a standard environment has reached an asymptote or evidences a slight constant slope, (b) day-to-day between-subject variance is constant, and (c) relative performance standings among subjects, as indicated by intertrial correlations, are constant from day to day. The importance of task stability has not been fully recognized in the development of previous batteries. Without stability, changes of the means during a repeated-measures experiment are not interpretable (Campbell & Stanley, 1963). In addition, stability ensures that the assumptions of repeated measures analysis of variance are met (Winer, 1971). Further, stability verifies the temporal generalizability (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) of subjects' scores. Lastly, stability ensures that what-is-being-measured does not change over time (Alvares & Hulin, 1972; Jones, 1980). As defined by Jones (1980), stability represents the properties which must be met for statistically and scientifically meaningful repeated-measures experiments.

In addition to stability, a test should be sensitive to environmental effects which are reflected in changes of the mean score associated with changes in treatment. Sensitivity to a change of the mean, it is pertinent to note, is enhanced by a large intertrial correlation (Winer, 1971). Figure 1 is a nomogram which shows the

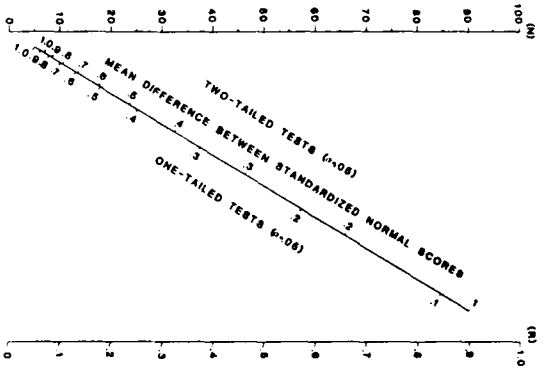


Figure 1. Nomogram showing the minimum statistically-significant difference ($p < .05$) between two trials of a repeated-measures experiment with sample size N and intertrial correlation R.

relationship between intertrial correlation (R), sample size (N), and the minimum statistically significant difference ($p < .05$) between standardized scores from two trials of a repeated-measures experiment. The nomogram is based on the equation given by Winer (1971) for testing differences between means of correlated observations. The figure shows, for example, that if one sets out to detect a mean change that exceeds .2 standard deviations (one tailed test), and if 20 subjects are available, then a task definition of .85 is required. Furthermore, the same significance level can be obtained for a mean difference of .3 standard deviations when N = 5, R = .90; or N = 33, R = .45; or N = 60, R = 0. This nomogram emphasizes the importance of intertrial correlation in the design of repeated measures experiments: a little intertrial correlation saves a lot of subjects.

Purpose

The primary purpose of this report is to present a description of the stability and other characteristics of 15 performance tasks which have been investigated as part of the PETER Program. A secondary purpose is to report progress on another 45 additional tests which are being studied. The goal of these presentations is to provide information useful to other investigators engaged in environmental research.

METHOD

The approach employed is to summarize information about candidate performance tasks in a tabular format. Twenty of the most relevant task characteristics were selected for presentation under two broad categories: (a) Background Information, and (b) Statistical Properties. Background Information included the ten characteristics defined in Table 1. Stability and

TABLE 1: DEFINITIONS OF BACKGROUND INFORMATION CHARACTERISTICS OF PERFORMANCE TASKS

CHARACTERISTIC		DEFINITION
1.	SOURCE REFERENCE	LITERATURE SOURCE DESCRIBING THE TASK
2.	PETER REFERENCE	REPORT ON THE TASK SUBJECTED TO PETER INVESTIGATION
3.	VALIDITIES	TYPES OF VALIDITIES TASK POSSESSES (CONTENT, CONSTRUCT, PREDICTIVE, FACE)
4.	VERIFICATION	CONTEXTS WHERE TASK HAS BEEN FOUND SENSITIVE
5.	INDIVIDUAL/ GROUP	TYPE OF ADMINISTRATION
6.	TEST MODE	APPARATUS REQUIRED (E.G. PAPER & PENCIL, T.V., AUDIO VIEWER, TIMER)
7.	TEST TIME IN SECONDS	TEST LENGTH IN SECONDS IN THE PETER EXPERIMENT
8.	SCORE	TYPE OF SCORE (I.E., HITS, % CORRECT, SLOPE, NUMBER ATTEMPTED, LATENCY)
9.	N	SAMPLE SIZE FOR WHICH DATA ARE AVAILABLE
10.	COMMENTS	CHARACTERISTICS WHICH DID NOT FALL CONVENIENTLY INTO OTHER CATEGORIES

sensitivity are described by the ten properties defined in Table 2. Most of the characteristics and properties listed in Tables 1 and 2 are easily understood. However, the "standardized reliability" (Table 2, Item 8) may require an explanation. It is the value, estimated by the Spearman-Brown formula (c.f. Winer, 1971), that the intertrial correlation would have had if the test had lasted three minutes. Standardized reliability is useful

TABLE 2: DEFINITIONS OF STATISTICAL PROPERTIES OF PERFORMANCE TASKS

PROPERTY		DEFINITION
1.	DAY $\bar{X}$ STABILIZES	DAY AT WHICH MEAN REACHES STABILITY
2.	$\bar{X}$	VALUE OF MEAN AT DAY STABILITY IS REACHED
3.	b_m	VALUE OF SLOPE OF SCORES DURING STABLE PERIOD
4.	DAY S.D. STABILIZES	DAY AT WHICH STANDARD DEVIATIONS BECOME STABLE
5.	S.D.	VALUE OF STABLE S.D.
6.	DAY R STABILIZES	DAY AT WHICH INTERTRIAL CORRELATION (R) STABILIZES
7.	TASK DEFINITION	VALUE OF R DURING STABLE PERIOD
8.	STANDARDIZED RELIABILITY	CALCULATED BY USING THE SPEARMAN-BROWN FORMULA USING A THREE MINUTE BASE (C.F., FIGURE 2)
9.	OVERALL STABILITY	DAY AT WHICH ALL FORMS OF STABILITY ARE PRESENT
10.	SENSITIVITY	DEGREE TO WHICH STANDARDIZED RELIABILITY EXCEEDS $r = .707$.

for comparing reliabilities of tests with different administration times. If such a comparison were made without regard to test administration time, then a test with a longer administration time would tend to be favored because reliability increases with test length. Figure 2 shows the tradeoff of test time and reliability, according to the Spearman-Brown formula. Standardized reliability allows comparisons of reliabilities of tests for any arbitrary (in this case, 3 minute) administration period.

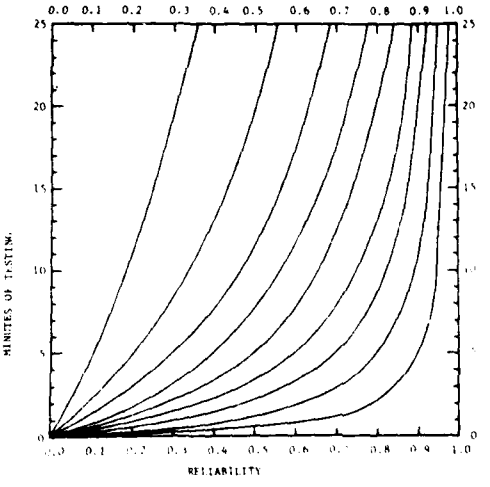


Figure 2. Tradeoff between intertrial correlation and test time.

RESULTS AND DISCUSSION

Fifteen completed appraisals of Performance Evaluation Tests for Environmental Research are summarized in Table 3. Some of the tests provide multiple scores. For example, the item recognition test yields a reaction time, slope and intercept. Because each score has its own properties and interpretations, the scores are represented by separate rows of Table 3. The first 10 columns of Table 3 list general characteristics (defined in Table 1) for each score. The remaining columns of Table 3 summarize the statistical results of the test assessments (defined in Table 2). Note that each score's mean stabilizes eventually (reaches constant slope). The mean ($\bar{X}$) at the day stability was attained, and the slope (b_m) that prevailed thereafter are listed in order that the mean on any particular stable day can be calculated. In contrast to the means, which usually required several sessions to stabilize, the standard deviations (S.D.) stabilized rapidly, usually during the first or second day of testing. At the other extreme, some of the intertrial correlation matrices never stabilized. Instead they exhibited superdiagonal form (Alvares & Hulin, 1972) throughout the 15 days of testing. However, most tests do provide stable intertrial correlations after several sessions of testing. Only a few of these tests have a creditable task definition. If it is required that the test predict at least 50% of its own variance in later sessions, then task definition would have to be in excess of .7. The extent to which this sensitivity criterion was met by each test is shown in the final column of Table 3. The penultimate column lists the days on which each test has stable means, S.D., and intertrial correlations. Considering both stability and sensitivity, six of the tests in Table 3 are recommended for inclusion in test batteries for environmental research using repeated measures: (1) Grammatical Reasoning, (2) Stroop, (3) Air Combat Maneuvering, (4) Code Substitution, (5) Arithmetic, and (6) Tapping.

Forty five additional tests are equally distributed among the three stages of appraisal: planning, data gathering, and analysis. More tests will be added to the program later. When the program was begun, it was assumed that 150 to 200 tests would be assessed to provide enough stable, sensitive tests to characterize human performance. Now, it is suspected that 100 tasks may suffice because several studies of tests representing presumably orthogonal factors have shown convergence (increased correlation) between the tests with extended practice (Kennedy, Bittner, & Jones, 1980; Jones, Kennedy, & Bittner, 1980; McCafferty, Bittner, & Carter, 1980).

It is anticipated that this is the first of many catalogues of Performance Evaluation Tests for Environmental Research. The tabular form of the catalogue is intended to provide useful information to environmental researchers in a succinct form. For instance, one may estimate the amount of distributed practice required for stability by multiplying "Administration Time" by "Day $\bar{X}$ Stabilizes". Furthermore, the catalogue provides information which may be used in conjunction with Figures 1 and 2 to plan sample size, testing time, and minimum detectable effects for repeated-measures experiments. However, the format of the catalogue is tentative. The authors encourage suggestions for a revised format to be used in future catalogues.

REFERENCES

1. Alvares, K. M., & Hulin, C. L. Two explanations of temporal changes in ability-skill relationships: A literature review and theoretical analysis. Human Factors, 1972, 14, 295-308.
2. Baddeley, A. D. A 3 min reasoning test based on grammatical transformation. Psychonomic Science, 1968, 10, 341-342.
3. Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research. Chicago: Rand McNally, 1966.
4. Carter, R. C., Kennedy, R. S., & Bittner, Jr., A. C. Grammatical reasoning: A stable performance yardstick, unpublished manuscript, 1980.
5. Carter, R. C., Kennedy, R. S., Bittner, Jr., A. C., & Krause, M. Item recognition as a performance evaluation test for environmental research. Proceedings of the 24th Annual Meeting of the Human Factors Society, 1980.
6. Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. The dependability of behavioral measurements. New York: John Wiley, 1972.
7. Damos, D. L., Kennedy, R. S., Bittner, Jr., A. C., & Harbeson, M. M. Effects of extended practice on dual-task training. Paper presented at the 87th Annual Convention of the American Psychological Association, 1979.
8. Damos, D. L., Kennedy, R. S., & Bittner, Jr., A. C. Development of a performance evaluation test for environmental research (PETER): Critical tracking test. Proceedings of the 50th Annual Scientific Meeting of the Aerospace Medical Association, 1979, 33-34.
9. Kennedy, R. S., Bittner, Jr., A. C., & Jones, M. B. Exploratory studies of tracking tasks. Unpublished manuscript, 1980.
10. Harbeson, M. M., Kennedy, R. S., & Bittner, Jr., A. C. A comparison of the Stroop test to other tasks for studies of environmental stress. Proceedings of the 12th Annual Meeting of the Human Factors Association of Canada, 1979, 21.1-21.9.
11. Jex, H. R., McDonnell, J. D., & Phatak, A. V. A "critical tracking task for manual control research. IEEE transactions on human factors in electronics. 1966, HFE-7: 138-145.
12. Jones, M. B. Stabilization and task definition in a performance test battery. (NBDL Monograph No. M-0001) New Orleans, LA: Naval Biodynamics Laboratory, 1980.

13. Jones, M. B., Kennedy, R. S., & Bittner, Jr., A. C. Video games and convergence or divergence with practice. Proceedings of the Seventh Psychology in the DOD Symposium, USAF Academy, Colorado Springs, CO, 16-18 April 1980.
14. Kennedy, R. S., & Bittner, Jr., A. C. The development of a Navy Performance Evaluation Test for Environmental Research (PETER). In L. T. Pope & D. Meister (Eds.), Productivity Enhancement: Personnel Performance Assessment in Navy Systems. Symposium presented at the Naval Personnel R & D Center, San Diego, October 1977, 393-408. (NTIS No. AD A045047)
15. Kennedy, R. S., & Bittner, Jr., A. C. Development of performance evaluation tests for environmental research (PETER): complex counting. Aviation, Space, and Environmental Medicine, 1980, 51, 142-144.
16. Kennedy, R. S., & Bittner, Jr., A. C. The utility of commercially available television-computer games for assessing performance and other applications. Proceedings of the 51st Annual Scientific Meeting of the Aerospace Medical Association, 1980.
17. Kennedy, R. S., Bittner, Jr., A. C., & Einbender, S. W. Development of performance evaluation tests for environmental research (PETER): trail making test. Unpublished manuscript, 1980.
18. Kennedy, R. S., Bittner, Jr., A. C., & Harbeson, M. M. An engineering approach to the standardization of performance evaluation tests for environmental research (PETER). Proceedings of the 11th Annual Conference of the Environmental Design Research Association (EDRA), Charleston, SC, 2-6 March, 1980.
19. Krause, M., & Kennedy, R. S. Performance evaluation tests for environmental research (PETER): Interference susceptibility test (IST). Proceedings of the 7th Psychology in the DOD Symposium. USAF Academy, Colorado Springs, CO, 1980.
20. McCafferty, D. B., Bittner, Jr., A. C., & Carter, R. C. Performance evaluation tests for environmental research (PETER): Auditory digit span, Proceedings of the 24th Annual Meeting of the Human Factors Society, 1980.
21. McCauley, M. E., Kennedy, R. S., Bittner, Jr., A. C. Development of performance evaluation tests for environmental research (PETER): time estimation test. Proceedings of the 23rd Annual Meeting of the Human Factors Society, Boston, MA, 513-517, October 1979.
22. Pepper, R. L., Kennedy, R. S., Bittner, Jr., A. C., & Wiker, S. F. Performance evaluation tests for environmental research (PETER): Code substitution test. Proceedings of the 7th Psychology in the DOD Symposium, USAF Academy, 1980.
23. Reitan, R. M., & Davison, L. A. Clinical neuropsychology: Current status and applications. New York: John Wiley, 1974.
24. Seales, D. M., Kennedy, R. S., & Bittner Jr., A. C. Development of performance evaluation tests for environmental research (PETER): arithmetic computation. Proceedings of the 23rd Annual Meeting of the Human Factors Society, 1979.
25. Sternberg, S. High speed scanning in human memory. Science, 1966, 153, 652-654.
26. Stroop, J. R. Studies of inference in serial verbal reactions. Journal of Experimental Psychology, 1935, 18, 643-662.
27. Underwood, B. J., Boruch, R. F., & Malmi, R. A. The composition of episodic memory. (ONR Contract No. N00014-76-C-0270) (NTIS No. AD A040696).
28. Wechsler, D. Measurement and appraisal of adult intelligence, Baltimore, MD: The Williams & Wilkins Co., 1939.
29. Welkind, I., & Sprug, J. Time research: 1,172 studies. Metuchen, NJ: Scarecrow Press, 1974.
30. Winer, B. J. Statistical principles in experimental design (2nd ed.). New York: McGraw-Hill, 1971.

TABLE 3: PRELIMINARY CATALOGUE OF 15 PERFORMANCE TASKS STUDIED BY THE PETER PARADIGM

	REFERENCE[a]																			
	SOURCE	PETER	VALIDITIES[b]	VERIFICATION[c]	INDIV./GROUP	TEST MODE[d]	TEST TIME (SEC)	SCORE[e]	N	COMMENTS[f]	DAY $\bar{X}$ STABILIZES	$\bar{X}$	b_m	DAY S.D. STABILIZES	S.D.	DAY R STABILIZES	TASK DEF'N.	STAN. RELIABILITY	DAY STABLE	SENSITIVITY[h]
GRAMMATICAL REASONING	2	4	2,4	2,3	G	1	60	NC	23	4	4	10	.37	1	5	5	.82	.93	5	++
ITEM RECOGNITION (a) SLOPE	25	5	2	1,3	I	3	900	S	21	2	2	45	0	2	60	–	–	.00	–	--
(b) RT	25	5	2	1,3	I	3	225	RT	21	2	4	700	0	2	20	3	.70	.65	4	–
COMPLEX COUNTING	15	15	4	3	G	4	900	PC	19	2,3	1	80	0	1	15	3	.85	.53	3	–
STROOP (a) BW WORDS	26	10	2	1,2,3	G	5	30	NA	19	3	7	53	0	1	8	4	.83	.92	7	++
(b) COLOR BLOCKS	26	10	2	1,2,3	G	5	30	NA	19	3	7	47	0	1	9	2	.83	.92	7	++
(c) COLOR WORDS	26	10	2	1,2,3	G	5	30	NA	19	3	7	40	0	1	10	4	.80	.90	7	++
AIR COMBAT MANEUVERING	16	16	4	4	I	2	1380	NC	15	2,5	4	13	.5	1	3	6	.93	.70	6	+
SLALOM	16	16	4	4	I	2	952	NC	15	2,5	4	79	2	1	13	7	.60	.22	7	--
DIGIT SPAN (a) FWD	28	20	1-4	1-4	G	6	1800	NC	9		3	20	.67	3	5	11	.68	.20	11	--
(b) BKWD	28	20	1-4	1-4	G	6	1800	NC	9		3	17	.67	3	8	–	–	.30	–	--
CODE SUBSTITUTION	28	22	1,2	1-4	G	1	240	NC	19	4	8	70	.8	8	10	8	.75	.70	8	+
ARITHMETIC	28	24	1,3,4	1-4	G	1	600	NC	18	4	4	36	0	1	18	1	.94	.85	4	++
TIME ESTIMATION	29	21	1,4	3	I	7	600	CE	19	5	1	8	0	1	3	–	–	.75	–	+
CRITICAL TRACKING	11	8	4	3	I	4	900	$\frac{1}{RT}$	18	1,2,5	4	5.5	.12	1	.7	10	.85	.65	10	–
DUAL CRITICAL TRACKING	11	7	4	2	I	4	900	$\frac{1}{RT}$	12	1,2,5	5	4.4	.07	4	.8	10	.76	.40	10	–
INTERFERENCE SUSCEPTIBILITY	27	19	2	4	I	3	600	PC	23	3	3	60	1.1	1	20	8	.71	.45	8	–
TRAIL MAKING	23	17	1,2	2	G	1	110	CT	18		5	110	0	2	22	2	.40	.50	5	–
TAPPING	23	17	1-4	1,2	G	1	36	CT	18		1	36	0	1	5	2	.85	.95	4	--

NOTES:

a. See References

b. Validities: 1-Content, 2-Construct, 3-Predictive, 4-Face

c. Verification: 1-Brain Damage, 2-Human Information Processing, 3-Environmental Change, 4-Factor identified by Factor Analysis

d. Test Mode: 1-Paper and Pencil, 2-T.V. Game, 3-Audioviewer, 4-Specialized Equipment, 5-Slides, 6-Verbal, 7-Stopwatch

e. Score: NC-Number Correct, S-Slope, RT-Reaction Time, PC-Percent Correct, NA-Number Attempted, CE-Algebraic sum of timing errors, CT-Completion Time

f. Comments: 1-Not Portable, 2-Possible electrical hazard in some environments, 3-limited number of forms available, 4-Computer programs available to generate forms, 5-Self scoring

g. Standardized Reliability: Estimate of what the reliability would have been if the test had lasted 3 minutes. Computed using the Spearman-Brown Formula (Winer, 1971)

h. Sensitivity: ++, $r \geq .8$; +, $.8 > r \geq .7$; -, $.7 > r \geq .35$; --, $r < .35$

PROCEEDINGS OF THE SEVENTH PSYCHOLOGY IN THE DOD SYMPOSIUM
USAF ACADEMY, COLORADO SPRINGS, CO 16-18 APRIL 1980

Performance Evaluation Tests for Environmental Research (PETER): Code Substitution Test

Ross L. Pepper
Naval Ocean Systems Center
Kailua, Hawaii

Robert S. Kennedy and Alvah C. Bittner, Jr.
Naval Aerospace Medical Research Laboratory Detachment
New Orleans, Louisiana 70189

Steven F. Wiker
Department of Industrial Engineering
University of Michigan, Ann Arbor 49107

Abstract

A Code Substitution Test was considered for inclusion in the Performance Evaluation Tests for Environmental Research (PETER) battery. The effects of repeated testing on code substitution performance was studied to determine reliability and stability of task performance. A single two minute testing trial per day was administered to a group of 19 subjects for 15 consecutive weekdays. In a second experiment, a four minute per day test was administered to 12 of the 19 original subjects for an additional 15 consecutive weekdays. Descriptive statistics are reported. Comparisons are made between these laboratory data and performances assessed at sea with repeated administration occurring within each day. The need for knowledge about task stability over repeated performance testing in exotic environments is discussed. The Code Substitution Test is recommended for inclusion in the PETER battery.

A research program is underway to evaluate tests of mental work for future use in studying adverse environments (Kennedy & Bittner, 1977). Each test is examined for stability as it is performed over periods of extended practice (15 days). Tests found to be suitably stable and to possess other characteristics (Kennedy, Bittner & Harbeson, 1980) are made part of a battery of Performance Evaluation Tests for Environmental Research (PETER). The present study reports the findings for a form of the Code Substitution (or Digit-Symbol) Test.

Otis is generally given credit for the initial development of a Digit-Symbol Test, and with Terman, the evolution of group intelligence testing around World War I (Wechsler, 1958). Wechsler (1958) included the Digit-Symbol Test in the original Wechsler-Bellevue (W-B) IQ Test. He felt this inclusion was required because it was one of the oldest and best established of all psychological tests. He felt that the Digit-Symbol Test measured both speed and power, and that both should be given weight in the evaluation of intelligence. He reported high correlations

The opinions are those of the authors and do not necessarily reflect those of the Department of the Navy.
This research was performed under Navy Work Unit No. MF58.524.002-5027.

between Digit-Symbol Test scores and total IQ scores (r = .673 for ages 20-34; r = .697 for ages 35-49 (see Wechsler, 1939 p. 136)). In describing the standardization of his test, Wechsler reported split-half coefficients ranging from r = .83 to r = .90 after correction for attenuation. However, it should be noted that his standardization procedure was not a conclusive demonstration of either reliability, stability, or validity. Correlations within and between the verbal and performance sub-tests indicated the measurement of common variation which could be either a common cluster of factors, correlated errors of measurement within days, or both. Hence, the consistency of the Digit-Symbol Test is not clear.

In addressing this issue, Derner, Aborn and Cantor (1950) rightly pointed out that the method of choice for determining the reliability of a measuring instrument is a test-retest technique. They then conducted a test-retest study to assess changes over 6 months, 4 weeks, and 1 week using normal adults (n=158). In all sub-tests, including Digit-Symbol, a learning effect was apparent. The overall WAIS reliability coefficients across test-retest intervals varied from r=.83 to r=.88 for the performance scale and Digit-Symbol was r=.80. This was the first substantial evidence that the Code Substitution Test has sufficient reliability to potentially reflect changes with environmental manipulations. It is noteworthy that except for the schizophrenic population employed by Ragin, all adult reliabilities on the Digit-Symbol test surveyed by Derner, et al. (1950) exceeded the mid .70's*. Hence, the body of literature suggests that the Digit-Symbol test has adequate simple test-retest reliability.

The stability of the Digit-Symbol test alone across extensive repeated testing or practice, has not been sufficiently established in previous research. The most relevant study was by Woodrow (1937) who compared the performance of high and low initial score performers on a variety of tests, including a Code Substitution Test. Testing was conducted daily for a 10 minute period for 39 days for one group (n = 56) and for 66 days for a second group (n = 82). The initial-final reliability coefficients for code substitution were r = .57 for the former, r = .59 for the latter. The ratio of initial and final group standard deviations were 1.57 and 1.64 respectively for the two groups, indicating that between subject differences increased slightly with practice, a finding that has been obtained elsewhere (Harbeson, Kennedy, & Bittner, 1979). The extent to which performance on a variety of tasks confounds findings is not known. Therefore, the primary purpose of the present effort was to study code substitution in the laboratory under baseline conditions over extended practice. A secondary purpose was to report the sensitivity of this test in a field study.

Experiment 1

Method

Subjects. Navy enlisted men (n=19) age 19-24 comprised the experimental group. These men were recruited, evaluated and employed in accordance with procedures described elsewhere (Thomas, Majewski, Ewing &

*Pristo (1978) has shown lower test-retest reliabilities ($\bar{r}$ = .20) in 40 children (IQ range 52-145) than expected.

Gilbert, 1978). These procedures meet or exceed prevailing national and international guidelines concerning human use in research. The subjects received extra compensation for volunteering and appeared motivated to perform. They were representative of the Navy population in size and intelligence but physically and mentally screened for hazardous duty environment research. They were under continuous medical supervision.

Apparatus. The Code Substitution test forms were derived after the concepts of Otis, where each day nine letters were randomly assigned a digit from one to nine. Fifteen alternate forms were computer generated following a general Monte Carlo algorithm: (a) the digit letter relationships were changed daily; (b) each letter appeared 10-15 times in a daily list of 135 items; and (c) each letter was nonrepeating. Figure 1 shows a layout of a sample test form.

Procedure. The subject's task was to follow the letter/number correspondence for a given day in assigning the appropriate letter below each number. Subjects were instructed to proceed rapidly and accurately throughout the list until told to stop. Each session in Experiment 1 lasted two minutes. The subjects were ordinarily tested in a group each workday morning for three weeks. Performance was scored according to number attempted, number correct, and rights minus wrongs. Group means, between subject standard deviations, and cross session reliabilities were calculated for each score. Analysis of variance (ANOVA) was conducted for days and subjects main effects.

Results

Only results for total-correct are reported here as the subjects made very few errors, (1 on the average/per subject/per day) and other scores (e.g. total attempted) were redundant. Figure 2 shows means and standard deviations for total-correct for nineteen subjects over 15 days. Mean performance is seen to improve throughout the study, although the trend becomes less pronounced after Day 8. Similarly, standard deviations increase but are relatively constant after Day 8. Figure 3 shows the cross session reliabilities for selected base days, the source of which is Table 1. Correlation traces (Bittner, 1979) show negative slopes for Base Days 1, 2, and 4. This trend is less evident in traces for Base Days 8, 10, and 12, suggestive of differential stabilization somewhere between Days 4 and 8. Task definition (Jones, 1980), the degree to which a test differentiates reliably between individuals, is greater than r = .75 subsequent to Day 8.

Experiment 2

Method

Subjects. Twelve of the 19 original subjects comprized the experimental group. Between the end of Experiment 1, and the beginning of Experiment 2, the other 7 subjects were transferred and were not available for testing.

Apparatus. The test forms were produced in the same way as in Experiment 1, with the exception that each day's test was twice as long (270 vs. 135 items).

Procedure. The procedure was the same as Experiment 1, except that the subjects were given 4 minutes rather than 2 minutes of testing each day. The testing period began 11 weeks after the conclusion of the first experiment, and continued for 15 consecutive workdays.

Results

Experiment 2 was conducted in an attempt at improving the magnitude of the correlation level by doubling testing time. Although only twelve of the original 19 subjects remained available for the retest, their means (Figure 5) were not statistically different from the original group ($p > .5$). The second study also was continued for fifteen days, and the means and standard deviations for these twelve subjects appear in Figure 6. While performance continued to improve over the period of the experiment, the change is slight but significant ($p < .01$). Not unexpectedly, the values are about twice those of the shorter test (cf. Figure 3). Correlations are level for all comparisons indicating task stabilization was manifested on Day 1 of Experiment 2. Task definition is better than with the shorter test but slightly less than predicted by a Spearman-Brown adjustment.

Experiment 3

Method

Subjects. Six U. S. Coastguardsmen were selected from the complement of the WPB 95 (White Patrol Boat) employed in this study.

Apparatus and Procedure. Testing materials and procedures were similar to those employed in Experiments 1 and 2 with the following exceptions: Testing was conducted hourly from 0800-1600 for four consecutive days. The testing compartment was located amidships, below decks. The first two days of testing were conducted dockside, with engines running. The second two days of testing occurred while the vessel steamed a double octagonal pattern seven miles southwest of Honolulu in the Molokai Channel, an area acknowledged for its turbulent sea condition. The testing commenced each sea day while the vessel steamed directly into the primary swell. Course changes of 45^{o} were made every half hour throughout the day, creating a systematically changing motion environment. (See Wiker & Pepper, 1978 for greater details of the testing conditions and a description of other task and subject variables assessed during this phase).

Results

Figure 7 shows performance on the Code Substitution Test for the six Coastguardmen exposed to mild seas in the Molokai channel. The data are plotted as scores per minute for the 16 dockside practice trials versus the 16 at sea data points. For comparability, the data from the first and second laboratory studies (Figures 4 and 5) have been replotted as a function of cumulative practice. Plotted in this way, 15 days of 2 minute laboratory trials can be compared to the first 15 hours of dockside testing.

The fit between the two studies for the first 30 minutes of practice is surprisingly good considering the known differences in the two experiments: (a) design - all performance massed in 4 days versus distributed over two 3-week periods 11 weeks apart; (b) test length - 2 and 4 minute trials were combined in the laboratory study versus two minute trials only in the field study; and (c) subjects - Navy versus Coastguardsmen. Secondly, the fit is also good during the sea trials with the exception of the second hour at sea where the poorest performance of all was obtained. This finding of performance degradation is concordant with the high motion sickness symptoms during this time frame (Wiker, Kennedy,

McCauley & Pepper, 1979). Moreover, because of the stability and differentiation of the laboratory version of the task and the close agreement between the two studies after the at sea decrement, the authors are inclined to consider this a real effect of motion on performance.

Discussion

The PETER Program is underway whereby psychological tests are being examined critically to determine their suitability for use in detecting performance degradation in novel environments (Kennedy & Bittner, 1977). The criteria against which tests are compared focus on stability and sensitivity. Stability is measured by examining the effects of extended practice on means, standard deviations and cross session reliabilities. Means are stable if they are level, asymptotic or exhibit constant slope. Standard deviations may be level or increase slightly with the mean. Cross session reliabilities are considered stable after they cease to change over sessions. In this study, qualify for the PETER battery. Means have constant slope after Day 8 of Experiment 1 and standard deviations are also level after that time. The reliabilities are moderate $\bar{r} > .75$ and stable after Day 8. Experiment 2 showed several things: (a) stability is still present 3 months later; (b) a test twice as long only improves reliability to an average of r=.80 while effectively doubling mean performance. This Code Substitution Test appears to be an excellent candidate for inclusion in PETER from the laboratory results.

The results of the sea trials in this study provide at the same time vindication and validation of the PETER paradigm. The laboratory task sufficiently differentiates subjects, and is stable, so that slight departures may be ascribed as due to environmental and not artifactual variables. The benefit of being able to compare real world performances at sea with those of a control group in a laboratory is also noteworthy. Both laboratory and environmental results recommended the use of the Code Substitution Test in PETER or other environmental batteries.

References

Bittner, Jr., A. C. Tests of differential stability. Proceedings of the 23rd Annual Meeting of the Human Factors Society, Boston, October, 1979.

Derner, G. F., Aborn, M. and Cantor, A. M. The reliability of the Wechsler-Bellevue Subtest and Scales. Journal of Consulting Psychology, 1950, 14, 172-179.

Harbeson, M. M. Kennedy, R. S. & Bittner, A. C., Jr. A comparison of the Stroop Test to other tasks for studies of environmental stress. Proceedings of the 12th Annual Meeting of the Human Factors Association of Canada, Bracebridge, Ontario, Canada, September, 1979.

Jones, M. B. Stabilization and task definition in a performance test battery. (NAMRL Monograph No. 27). Pensacola, FL: U.S. Naval Aerospace Medical Research Laboratory, 1980.

Kennedy, R. S. & Bittner, A. C., Jr. The development of a Navy Performance Evaluation Test for Environmental Research (PETER). In, Productivity Enhancement: Personnel Performance Assessment in

Navy Systems. Symposium presented at the Naval Personnel Research and Development Center, San Diego, CA, 12-14 October, 1977. (NTIS No. AD-056047).

Kennedy, R. S., Bittner, A. C., Jr., & Harbeson, M. M. An engineering approach to the standardization of Performance Evaluation Tests for Environmental Research (PETER). Proceedings of the 11th Annual Conference of the Environmental Design Research Association, Charleston, S.C., March, 1980.

Pristo, L. J. Comparing WAIS and WISC-R Scores. Psychological Reports, 1978, 42, 515-518.

Thomas, D. J., Majewski, P. L., Ewing, C. L. & Gilbert, N. S. Medical Qualification Procedures for Hazardousduty Aeromedical Research. (Conference Proceedings No. 231, A3, pp. 1-13, 1978) London: AGARD, 1977.

Wechsler, D. Measurement of Adult Intelligence (1st ed.). Baltimore: Williams & Wilkins Co., 1939.

Wechsler, D. The Measurement and Appraisal of Adult Intelligence. Baltimore: The Williams & Wilkins Co., 1958.

Wiker, S. F. & Pepper, R. L. Change in crew performance, physiology and affective state due to motions aboard a small monohull vessel; a preliminary study. Coast Guard Technical Report No. CG-D-75-78, 1978.

Wiker, S. F., Kennedy, R. S., McCauley, M. E., & Pepper, R. L. Susceptibility to seasickness: Influence of hull design and steaming direction. Aviation, Space & Environmental Medicine, 1979, 50, 1046-1051.

Woodrow, H. Factors in improvement with practice. The Journal of Psychology, 1937, 7, 55-70.

Table

Table 1
Experiment 1: Comparison of Reliabilities for Total Correct Over 15 Days (n=18)

Days	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	.60	.50	.23	.68	.76	.87	.54	.65	.39	.27	.63	.30	.65	.53
2		.74	.53	.72	.60	.62	.68	.50	.61	.41	.51	.53	.58	.53
3			.77	.81	.62	.49	.81	.58	.74	.57	.62	.66	.62	.67
4				.73	.51	.29	.78	.68	.82	.67	.55	.70	.54	.68
5					.86	.73	.87	.77	.80	.70	.71	.75	.84	.78
6						.78	.78	.57	.78	.60	.75	.67	.86	.74
7							59	.65	.56	.44	.73	.54	.72	.65
8								.67	.90	.77	.68	.74	.65	.69
9									.72	.71	.73	.77	.68	.76
10										.84	.74	.86	.68	.74
11											.71	.85	.58	.76
12												.78	.75	.85
13													.77	.87
14														.85

Figures

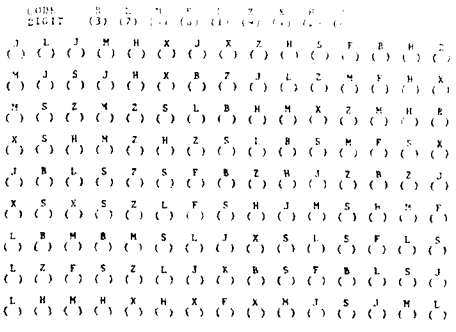


Figure 1. Coding Test Sample Form (Day 1, Experiment 1)

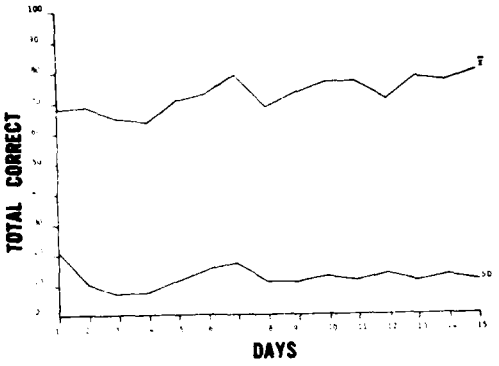


Figure 2. Experiment 1: Means and Standard Deviations for Total Correct Over 15 Days (n=19).

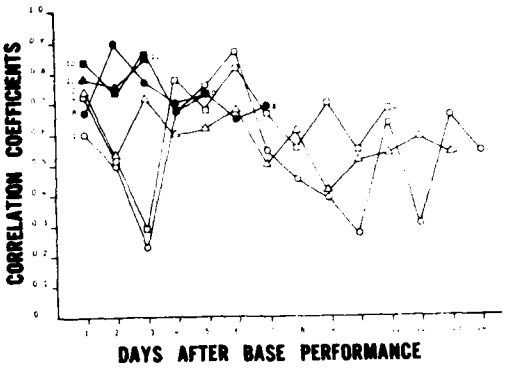


Figure 3. Experiment 1: Correlations Between Selected Base Days (1, 2, 4, 8, 10, 12) and Those Following For Total Correct Over 15 Days (n=19).

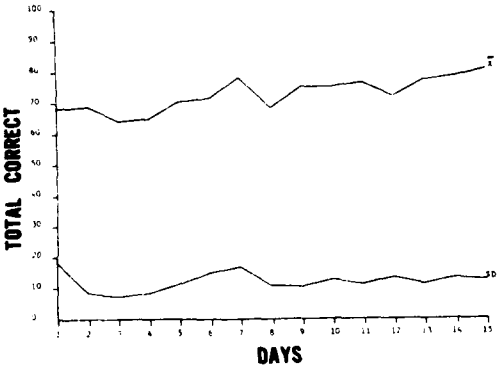


Figure 4. Experiment 1: Means and Standard Deviations for Total Correct Over 15 Days (n=12).

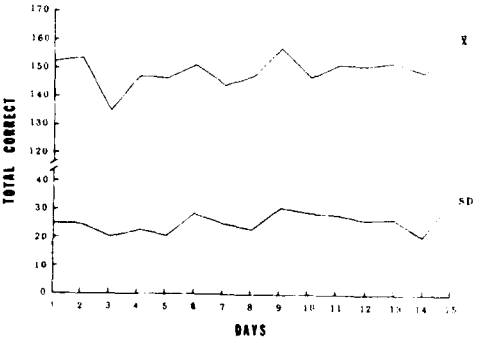


Figure 5. Experiment 2: Means and Standard Deviations for Total Correct Over 15 Days (n=12).

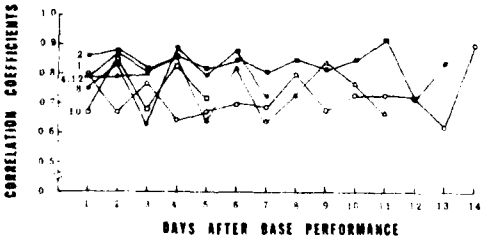


Figure 6. Experiment 2: Correlations Between Selected Base Days (1, 2, 4, 8, 10, 12) and Those Following for Total Correct Over 15 Days (n=12).

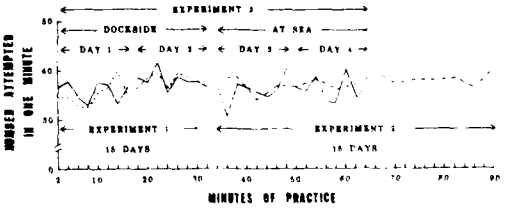


Figure 7. Experiments 1, 2, and 3: Mean Number Attempted in 1 Minute.

Proceedings of the 12th Annual Meeting of the Human Factors Association of Canada
Bracebridge, Ontario, Canada, September, 1979

A COMPARISON OF THE STROOP TEST TO OTHER TASKS FOR STUDIES OF ENVIRONMENTAL STRESS

Mary M. Harbeson, Robert S. Kennedy and Alvah C. Bittner, Jr.

Naval Aerospace Medical Research Laboratory Detachment
P. O. Box 29407, New Orleans, Louisiana 70189

ABSTRACT

A program is underway to standardize a battery of Performance Evaluation Tests for Environmental Research (PETER). The purpose of the program is to develop a test battery which will measure the effects of extended exposure to unusual environments (e.g. ship motion and vibration) on the performance of U. S. Navy personnel. Tasks which meet one or more of the following criteria are being examined: sensitivity to unusual environments, diagnostic capability for brain damage, or the ability to measure some aspect of information processing. The strategy for developing PETER has been to administer each task for 15 consecutive work days to the same group of 20 men who serve as volunteer subjects, and to examine the stability of the means, variances and reliabilities. These statistics thus become specifications which may be employed to evaluate and compare the suitability of tasks for inclusion in a test battery. This report focuses chiefly on the Stroop Test, and describes our approach in detail. The Stroop specifications are compared with "good" and "bad" tasks from our recent experiments. The tests used for comparison are: complex counting, critical tracking, time estimation, arithmetic and air combat maneuvering. Examples of tests which are unsuitable because of failure to meet only one of the three criteria are shown. The importance of the stability of the reliability, heretofore ignored in performance test battery construction is discussed.

INTRODUCTION

PETER Paradigm

An experimental program for the development of Performance Evaluation Tests for Environmental Research (PETER) is currently underway at the Naval Aerospace Medical Research Laboratory (NAMRL) (Kennedy & Bittner, 1977, 1978). The purpose of the program is to develop a test battery to determine if human performance is disrupted by the unusual environmental conditions experienced by Navy personnel (e.g., ship motion and vibration) over extended exposures. The program is designed to resemble an engineering test and evaluation program, since each test or element is subjected to an analysis of its performance specifications.

Specifically, baseline measures of performance are obtained in a series of tests administered for 15 consecutive weekdays to the same group of subjects. Three statistical criteria are being considered in the evaluation of the suitability of a test for use in unusual environments, viz. means, variances, and cross session reliabilities. Whereas stable means are intuitively desirable for the study of environmental effects, we feel that other approaches based upon reliability are more relevant. For example, when subjects serve as their own control, task reliability can sharpen the Student's t Test by reducing the standard error which appears as the denominator in equation (1).

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2 + S_2^2 - 2r_{12} S_1 S_2}{N}}} \qquad (1)$$

In particular, it may be seen that as the correlation approaches unity and the variances remain equivalent, the denominator of (1) will approach zero. Other statistical considerations also direct attention to reliability as a criterion. Simple repeated measures analysis of data, in particular, require stability of reliabilities across trials (Jones, 1979; Bittner, 1979). Hence, the strategy for building PETER, in addition to monitoring changes in the means and variance has been to focus on the stability of the reliabilities of a test over many sessions. Each of these statistical criteria warrants separate discussion.

Means. It is felt that there are three criteria for mean stability: (1) Plateaus are most desirable but they occur infrequently (cf. e.g., Kennedy & Bittner, 1978); (2) Asymptotic means are acceptable but are not always obtained even when practice is extended (Bradley, 1969); and (3) Jones (1979) has suggested that a slow regular, linear increase over sessions also reflects stability.

Standard deviations. Whereas the within-subject variance can be expected to decrease with practice, it is the between subject variances which are listed in equation (1). These between-subject variances may be considered stable when they are constant. In addition, as the means increase, it is possible that standard deviations will also increase with practice. (Jones, 1971, p 109). Standard deviation stability is considered to be present in this latter case if there is a concordant stabilization in the means and correlations.

Correlations. Since at least the time of Perl (1934), it has been known that during the acquisition phases of practice, the cross session reliabilities can be expected to change. This change takes the general form of a decrease along any row in the correlation matrix beginning with the superdiagonal, and has been referred to as Simplex form (Humphreys, 1960; Jones, 1969). It has been inferred that when these correlations cease to change within the matrix, then the task has differentially stabilized (Jones, 1969). We concur with this criterion and employ graphical analysis to determine where and if stabilization is obtained. The level at which a task differentiates subjects after it has stabilized is also an important

* This research was performed under Navy Contract No. MF58.524.002-1027. The opinions are those of the authors, and do not necessarily reflect those of the Department of the Navy.

factor involving the cross session reliabilities. High correlations obviously are most desirable and r=.707 is considered to be a lower limit for inclusion in PETER.

Task selection criteria. Candidate subtasks must meet one or more of the following criteria in order to be evaluated for inclusion in PETER: sensitivity to unusual environments; neuropsychological diagnostic capability; the ability to measure some aspect of information processing; and practical (e.g., cost) considerations (Kennedy & Bittner, 1977). The Stroop Test (Stroop, 1935) has been reviewed extensively (Jensen & Rohwer, 1966; Dyer, 1973). It was chosen for study since it met all our criteria for test selection.

The Stroop Test

Background. The Stroop test has been applied in a wide variety of investigations and will be used as an example of the analysis applied in the PETER program. It has been used as a measure of psychological stress in environmental studies (Reilly & Cameron, 1968; Biersner & Cameron, 1970; Schilling, Werts & Schandelmeier, 1976; Allan, Gibson & Green, 1979). Further, it has been shown to be sensitive to age, drugs, psychiatric disturbance and organic brain damage (Jensen & Rohwer, 1966; Comalli, Wapner & Werner, 1962; Dyer, 1973). It has frequently been used in the study of information processing functions (Stroop, 1935, 1938; Jensen & Rohwer, 1966; Dyer, 1973; Rose, 1974; Williams, 1977). In addition, the Stroop test has many attractive practical features. It can be group administered, takes very little time, and the apparatus is simple, economical and portable.

The Stroop test is reported to provide measures of individual differences on three factors: a speed factor, color-naming facility, and (of greatest interest to investigators) interference proneness (Jensen & Rohwer, 1966). The interference score, or "Stroop phenomenon", is the increase in reaction time between naming a color and naming the color of words printed in incompatible colors. This score is described as an index of susceptibility to mild stress (Thurstone & Mellinger, 1953, cited in Jensen, 1966; Sarmany, 1977) or the ability to resist distraction (Comalli, Wapner & Werner, 1962) although the generality of this finding has yet to be demonstrated. The psychological characteristics of the Stroop appear to be primarily in the cognitive realm. (Dyer, 1973; Golden, Marsella & Golden, 1975; Jensen & Rohwer, 1966; Sarmany, 1977). Stated differently, individual differences in the "Stroop phenomenon" are most likely related to differences in perceptual style.

In summary, there is sufficient research to suggest that performances on the Stroop Test tap an important faculty of an individual. Moreover, it can be inferred that this faculty is related to the work that Naval personnel perform during the course of their motion exposures, at sea and in flight. Regardless of whether the faculty is called interference proneness or stress susceptibility, it remains to be determined whether this faculty is an enduring aspect of an individual.

Alternate forms. Many versions of the Stroop Test are available but most use the following three conditions: (a) black and white words (BW) - color names written in black and white; (b) color blocks (CB) - blocks of color (usually red, blue and green) contained in a single, specified shape; and (c) a color-word condition (CW) - color names written in incompatible colors (e.g., the word "red" printed in blue). In previous research, methods of administration and scoring have varied but the interference effect has still been obtained. Subjects have been required to make verbal responses or manual responses (Flowers & Stoup, 1977; Jensen & Rohwer, 1966), such as key pressing (Keele, 1972) or card sorting (Stroop, 1938). Individual and group administration have been employed (Golden, 1975; Jensen & Rohwer, 1966). Mode of presentation and arrangement of stimuli, as well as number of colors (Golden, 1974; Jensen & Rohwer, 1966; Williams, 1977) have also been varied. Numerous (>20) scoring methods have been developed by the many investigators who have employed the Stroop Test (Jensen, 1965).

Adaption for PETER. In order to adapt the Stroop Test for environmental testing, group administration with manual responses was selected, and slides were used for presentation. The arrangement of stimulus material, conditions, colors, and method of scoring were those used most commonly in other studies (Jensen & Rohwer, 1966; Jensen, 1965; Dyer, 1973).

Other Tests

The Stroop Test results were compared with those obtained on five other tasks which have also been studied for inclusion in PETER: complex counting, critical tracking, time estimation, arithmetic computation, and air combat maneuvering. All tests were administered and analyzed according to the PETER paradigm. In the complex counting test (Kennedy & Bittner, 1979), subjects listened to three tones played simultaneously, and were required to keep track of every fourth low and medium tone. For the critical tracking test, (Damos, Kennedy & Bittner, 1979) the apparatus was a replication of that used by Jex, McDonnell & Phatak, 1966. In the time estimation test (McCauley, Kennedy & Bittner, 1979), subjects produced time intervals by verbal request. The arithmetic test (Seales, Kennedy & Bittner, 1979) was comprised of a paper and pencil presentation of simple arithmetic operations. The air combat maneuvering test (Jones, Kennedy & Bittner, 1979) was an adaptation of an Atari® Video Game (Atari, 1977) in which the subjects attempted to hit a moving drone with a missile.

Purpose

The purpose of this study was to determine the suitability of a group administered form of the Stroop Test by examining the effects of many sessions on the reliability, variability, and mean performance of three basic scores (BW, CB and CW) and two derived scores (BW-CB and CB-CW). The Stroop "specifications" were then compared to those of five other tests previously studied in the PETER program: complex counting, critical tracking, time estimation, arithmetic computation, and air combat maneuvering.

METHOD

Subjects

The subjects were a group of 19 Navy enlisted men, ages 19 to 24, who had served as volunteer subjects in several biodynamics studies since induction into the Navy (approximately 18 months prior to the testing). To qualify for this medical research program, they had to be equal or above the norms for Navy enlisted personnel in physical health, mental health and intelligence. All volunteer subjects were recruited, evaluated and employed in accordance with procedures specified in Secretary of the Navy Instruction 3900.39 and Bureau of Medicine and Surgery Instruction 3900.6 which are based upon voluntary informed consent, and meet the provisions of prevailing national and international guidelines. A

description of the subject selection procedure is given by Thomas, Majewski, Ewing and Gilbert (1977).

Apparatus

Slides (35 mm) were used to present the stimulus material for the three conditions, BW, CB and CW. The items on each slide were arranged in a 10 X 10 matrix of evenly spaced rows and columns. The colors red, blue and green were used. Rectangles of color were used for the CB slide. Items on all cards were in random order. There were two alternate forms for each condition. The slides were presented by means of a Kodak Carousel Projector (750H), and projected on a 1.45M X 1.32M movie screen which was placed approximately 3 meters from the subjects who were seated in armchair desks. Subjects responded by pushing buttons labeled, left to right, "R" for red, "B" for blue, and "G" for green, which were located on small switch boxes that were placed on each desk top. Subjects responses were automatically recorded on instrument chart paper. A Kronos stopwatch was used to regulate both the slide-viewing time and the inter-trial interval. The arrangement of the apparatus provided for testing in groups of four.

Procedure

The two alternate forms for each condition were arranged in eight possible combinations. A different order of presentation was used each day for eight days and seven of the combinations were repeated, one for each day, for the last seven days of testing. In the initial experimental session, after extensive practice on the use of the response keys, the subjects were instructed to begin responding to each slide immediately after it appeared on the screen. Instructions to the subjects for each of the 3 slides in the order in which they appeared were: (a) BW - to push the buttons corresponding to the color names as they appeared; (b) CB - to push the buttons corresponding to the color blocks as they appeared; (c) CW - to push the button corresponding to the color that each word was written in, regardless of the color that the word described. Each of the slides remained on the screen for 30 seconds and the inter-trial interval was 5 seconds. The same procedure, with the exception or abbreviation of instructions, was followed on subsequent testing days. The response measure was the number of responses in 30 seconds for each condition.

RESULTS

Figure 1 shows mean performance for the three directly measured scores (BW, CB & CW) and the two derived scores (BW-CB and CB-CW). The overall impression for the directly measured scores is of learning curves which are near asymptote after Day 10. The two derived scores, CB-CW and BW-CB, appear to approach an asymptote subsequent to Day 6. Mean responses for BW and CB were greater than CW throughout the test. Standard deviations for the three direct and two derived scores are given in Figure 2. It may be seen that the direct scores appear relatively stable and appear to covary with the means in Figure 1. In other words, there is slightly more variablity as the mean responses increase, following the general rule described by Jones (1972). Standard deviations for the two derived scores appear nearly level. A two-way analysis of variance, repeated measures design, showed significant days (practice) and subjects effects for all scores ($p < 10^{-3}$).

Tables 1 through 5 contain the correlations (reliabilities) over 15 days for the direct and derived Stroop scores. Figures 3 through 7 were drawn from these tables and show reliability "traces" for selected Base Days (1, 2, 4, 9 and 13) for the five scores. Trace plots were made of the correlations of each base day with those following, i.e., (Base Day 1 with 2, 1 with 3, 1 with 4 ..., 1 with 15; Base Day 2 with 3, 2 with 4, 2 with 5 ... 2 with 15, etc.) A fuller description of the construction and intrepretation of this type of plot is given elsewhere (Bittner, 1979). Examining these figures, it may be seen that BW (Figure 3), CB (Figure 4) and CW (Figure 5), reliabilities are relatively high after the early base days. For example, on BW the correlations of Base Day 1 to the days after base performances ranges between $r = .5$ and $r = .7$, while the correlation of Base Day 9 to subsequent days is of the order of $r = .9$. CB (Figure 4) proved to be most reliable with virtually all correlations of base days to subsequent days ranging from about $r = .75$ to $r = .96$. BW was more reliable than CW for early days after base performance, but there is a more pronounced decline in reliability for CW. The derived scores, BW-CB (Figure 6) and CB-CW (Figure 7) proved to be relatively unreliable, mutually ranging from the high of $r = .59$ to zero.

The results of the five tasks which were compared to the Stroop Test are summarized in Table 6. The means, standard deviations and correlations for selected days are shown in Tables 8 through 17.

DISCUSSION

From graphical analyses of the basic scores (BW, CB, and CW), it is apparent that these means and standard deviations are virtually stable after the initial base day's practice. In general it would also appear that a relatively stable and satisfactory level of reliability is available for all three of these measures subsequent to the early base days' practice. The means and standard deviations of the derived scores (BW-CB and CB-CW) (Figures 6 & 7) also show invariant behaviors over 15 sessions, but the reliabilities were extremely low.

It is possible that the reliability of the derived scores could be increased by making some changes in the administration of the test. A longer session, each performance day, could be expected to raise reliabilities, perhaps with greater spacing between the BW, CB and CW tasks. It is also possible that the amount of interference could be increased by changing the test in other ways. It has been found in previous studies that when motor rather than verbal responses are required, the color naming response is greater than the reading response (Flowers & Stoup, 1977; Keele, 1972; Stroop, 1938). In the present study, the response keys were marked with letters, thus combining reading and manual responses initially, although the letter-color relationships were considerably over-learned. Perhaps a purer measure of interference, and greater reliabilities in the derived scores, could be obtained by changing the response requirement to verbal rather than manual. This modification would limit the usefulness of the test for environmental test purposes; however, since group administration is of considerable practical importance (Kennedy & Bittner, 1977).

Regardless of whether or not the reliability of the derived scores could be improved by changing the testing procedure, the important point is, that the problem could not have been identified without examining all three statistical criteria. To further illustrate the importance of this type of analysis, and to demonstrate the possible combinations of means,

standard deviations and correlations, the Stroop results were compared to five other tasks which have been studied in the PETER program.

The analyses of the five other tests (Figures 8-17) follows the same paradigm as shown for the five Stroop scores. These five tests were selected from over 50 experiments since they contained examples of our major findings to date concerning task stabilization. Stabilities of means, standard deviations and cross session reliabilities were judged according to the criteria listed previously. These judgments are summarized in Table 6. It is our opinion that the most important finding is this table in that means alone (even means + standard deviations) are inadequate for determining stability. This finding achieves greater importance when viewed in connection with the scientific literature which reports performances in exotic environments. It is quite possible that no experiment has ever been performed in an unusual environment whereby adequate task stabilization was obtained in the pretest condition.

In summary, a group form of the Stroop Test was administered according to the PETER paradigm. Means, standard deviations and correlations were examined and compared with those from five other tasks. It was concluded that the three basic scores of the Stroop Test (BW, CB and CW) appear to be acceptable for inclusion in PETER. However, the lack of derived score reliabilities suggests that neither of these scores in their present form characterize a sufficiently stable faculty of mental work to be useful in the study of unusual environments.

REFERENCES

Allan, J. R., Gibson, T. M. & Green, R. G. Effect of induced cyclic changes of deep body temperature on task performances. Aviation, Space, and Environmental Medicine, 1979, 50, 585-589.

Atari, Inc. Combat game program instructions. Sunnyvale, California: Atari, Inc., Consumer Division, 1977. (C011402-01).

Biersner, R. J. & Cameron, B. J. Cognitive performance during a 1000-foot helium dive. Aerospace Medicine, 1970, 41, 918-920.

Bittner Jr., A. C. Statistical tests for differential stability. Proceedings of the 23rd Annual Meeting of The Human Factors Society, Boston, October, 1979 (in press).

Bradley, J. V. Practice to an asymptote. Journal of Motor Behavior, 1969, 1, 283-293.

Comalli, B. E., Wapner, S. & Werner, H. Interference effects of Stroop Color-Word Test in childhood, adulthood and aging. Journal of Genetic Psychology, 1962, 100, 57-83.

Damos, D. L., Kennedy, R. S. & Bittner, Jr., A. C. Development of Performance Evaluation Tests for Environmental Research (PETER): Critical tracking test. Proceedings of the 50th Annual Meeting of the Aerospace Medical Association, Washington, D.C., May, 1979. (AD A066719)

Dyer, F. N. The Stroop phenomenon and its use in the study of perceptual, cognitive, and response processes. Memory and Cognition, 1973, 1, 106-120.

Flowers, J. H. & Stoup, C.M. Selective attention between words, shapes and colors in speeded classification and vocalization tasks. Memory and Cognition, 1977, 5, 299-307.

Golden, C. J. Effect of differing number of colors on the Stroop color and word test. Perceptual and Motor Skills, 1974, 39, 50.

Golden, C. J. A group form of the Stroop color and word test. Journal of Personality Assessment, 1975, 39, 386-388.

Golden, C. J., Marsella, A. J. & Golden, E. E. Personality correlates of the Stroop color and word test: more negative results. Perceptual and Motor Skills, 1975, 41, 599-602.

Humphries, L. G. Investigation of the simplex, Psychometrika, 1960, 4, 313-323.

Jensen, A. R. Scoring the Stroop Test. Acta Psychologica, 1965, 24, 398-408.

Jensen, A. R. & Rohwer, W. D. The Stroop Color-Word Test: A review. Acta Psychologica, 1966, 25, 36-93.

Jex, H. R., McDonnell, J. D. & Phatak, A. V. A "critical" tracking task for manual control research. IEEE Transactions on Human Factors in Electronics, 1966, HFE-7, 138-145.

Jones, M. B. Individual differences. In R. N. Singer (Ed.). The Psychomotor Domain. Philadelphia: Lea and Fabinger, 1972.

Jones, M. B. Differential processes in acquisition, In E. A. Bilodea and I. McD. Bilodeau (Eds.), Principles of skill acquisition. New York: Academic Press, 1969.

Jones, M. B. Stabilization and task definition in a performance test battery. Pennsylvania State University College of Medicine, Final Report on Contract N0023-79-M-5089, May 1979.

Jones, M. B., Kennedy, R. S. & Bittner, Jr., A. C. A video game for performance testing. Paper presented at the Rocky Mountain Psychological Association Annual Meeting, Los Vegas, NV, May 1979.

Kennedy, R. S. & Bittner Jr., A. C. The development of a Navy Performance Evaluation Test for Environmental Research (PETER). In Productivity Enhancement: Personnel Performance Assessment in Navy Systems, Naval Personnel Research and Development Center, San Diego, CA 12-14, October 1977. (AD A056047)

Kennedy, R. S. & Bittner Jr., A. C. Progress in the analysis of a Performance Evaluation Test for Environmental Research (PETER). Proceedings of the 22nd Annual Meeting of the Human Factors Society, Detroit, MI, October 1978. (AD A060676)

Kennedy, R. S. & Bittner, Jr., A. C. Development of Performance Evaluations Tests for Environmental Research (PETER): Complex counting test. Journal of Aviation Space and Environmental Medicine. (in press)

Keele, S. W. Attention demands of memory retrieval. Journal of Experimental Psychology, 1972, 93, 245-248.

McCauley M. E., Kennedy, R. S. & Bittner, Jr., A. C. Development of Performance Evaluation Tests for Environmental Research (PETER): Time estimation test. Proceedings of the 23rd Annual Meeting of the Human Factors Society, Boston, October, 1979 (in press).

Perl, R. E. An application of Thurstones' method of factor analysis to practice series. Journal of Geneal Psychology. 1934, 11, 209-212.

Reilly, R. E. & Cameron, B. J. An integrated measurement system of the study of human performance in the underwater environment. Falls Church, VA: Bio-Technology, Inc. 1968.

Rose, A. M. Human Information Processing: An Assessment and Research Battery. Doctoral Dissertation, Ann Arbor, MI: University of Michigan, 1974, (also published as AFOSR-PR-74-1372). AD-785-411.

Sarmany, I. Different performance in Stroop's interference test from the aspect of personality and sex. Studia Psychologica, 1977, 19, 60-67.

Schilling, C. W., Werts, M. R. & Schandelmeier, N. R. (Eds.) The Underwater Handbook: A Guide to Physiology and Performance for the Engineer. New York: Plenum Press, 1976.

Seales, D. M., Kennedy, R. S. & Bittner, Jr., A. C. Development of Performance Evaluation Tests for Environmental Research (PETER): Arithmetic computation. Proceedings of the 23rd Annual Meeting of the Human Factors Society, Boston, October, 1979 (in press).

Stroop, J. R. Studies of interference in serial verbal reactions. Journal of Experimental Psychology, 1935, 18, 643-662.

Stroop, J. R. Factors affecting speed in serial verbal reactions. Psychological Monographs, 1938, 50, 38-48.

Thomas, D. J., Majewski, P. L., Ewing, C. L. & Gilbert, N. S. Medical Qualification Procedures for Hazardous-duty Aeromedical Research. (Conference Proceedings No. 231, A3, pp. 1-13, 1978) London: AGARD, 1977.

Williams, E. The effects of amount of information in the Stroop color word test. Perception and Psychophysics, 1977, 22, 463-470. (a)

TABLES

Table 1
Black and White Word
Reliabilities Over 15 Days (n=19)

Days	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	.67*	.70	.61	.80	.54	.68	.59	.55	.52	.65	.61	.64	.56	.69
2		.78	.64	.78	.75	.77	.75	.64	.56	.69	.76	.66	.63	.73
3			.80	.84	.89	.93	.86	.85	.82	.89	.90	.77	.75	.88
4				.86	.75	.85	.80	.90	.77	.85	.84	.89	.92	.73
5					.88	.89	.87	.89	.80	.92	.88	.81	.79	.87
6						.87	.84	.86	.84	.85	.86	.75	.78	.89
7							.93	.85	.83	.87	.91	.85	.80	.81
8								.68	.82	.85	.81	.74	.72	.80
9									.91	.93	.88	.87	.83	.86
10										.88	.89	.79	.71	.79
11											.86	.85	.73	.86
12												.78	.84	.84
13													.77	.73
14														.70

*r = .46 for p .05
r = .58 for p .01

Table 2
Color Block Reliabilities
Over 15 Days (n=19)

Days	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	.81*	.82	.78	.77	.78	.77	.84	.78	.79	.78	.82	.77	.81	.76
2		.96	.82	.91	.87	.93	.88	.91	.83	.95	.85	.83	.90	.77
3			.89	.91	.85	.93	.83	.86	.80	.91	.84	.86	.88	.74
4				.83	.77	.83	.73	.80	.79	.83	.76	.90	.82	.64
5					.96	.95	.90	.95	.87	.92	.90	.91	.95	.81
6						.91	.92	.94	.95	.91	.91	.89	.93	.86
7							.87	.89	.85	.97	.86	.84	.91	.76
8								.92	.87	.88	.92	.86	.92	.87
9									.93	.91	.90	.90	.94	.86
10										.88	.87	.87	.89	.84
11											.85	.87	.89	.81
12												.89	.91	.77
13													.90	.81
14														.82

*r = .46 for p .05
r = .58 for p .01

Table 3
Color Word Reliabilities
Over 15 Days (n=19)

Days	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	.82*	.76	.56	.66	.68	.43	.61	.32	.53	.56	.45	.43	.62	.57
2		.88	.65	.73	.78	.68	.78	.58	.70	.74	.71	.70	.74	.73
3			.79	.89	.86	.79	.91	.76	.81	.83	.80	.76	.91	.86
4				.74	.79	.76	.86	.80	.77	.72	.79	.76	.82	.76
5					.88	.86	.88	.76	.76	.87	.79	.83	.94	.88
6						.91	.90	.83	.74	.82	.87	.83	.87	.84
7							.89	.92	.78	.89	.87	.88	.86	.85
8								.89	.89	.90	.94	.87	.91	.92
9									.81	.84	.92	.85	.85	.86
10										.89	.89	.81	.83	.88
11											.86	.84	.84	.92
12												.90	.84	.91
13													.85	.88
14														.88

*r = .46 for p .05
r = .58 for p .01

Table 4
BW-CB Reliabilities
Over 15 Days (n=19)

Days	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	.30*	.26	.34	.19	.36	.39	.08	.53	.33	.55	.11	.51	.37	-.07
2		-.11	.18	.03	.28	.10	.35	.23	.07	.00	-.05	.24	.03	-.01
3			.47	.16	.36	.49	-.03	.04	.19	.44	.49	.24	-.12	-.11
4				.11	.37	.24	-.03	.31	.25	.27	.40	.50	.14	-.27
5					-.02	.51	.13	.32	-.29	.15	.20	.19	-.01	-.66
6						.42	.45	.59	.59	.20	.59	.58	.58	-.11
7							.29	.53	.18	.49	.24	.44	.10	-.33
8								.54	.32	.13	.56	.25	.24	-.26
9									.46	.49	.44	.50	.54	-.24
10										.40	.30	.52	.37	.07
11											-.00	.16	.06	.09
12												.36	.30	-.31
13													.36	-.38
14														-.10

*r = .44 for p .05
r = .58 for p .01

Table 5
CB-CW Reliabilities
Over 15 Days (n=19)

Days	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	.55*	.50	.69	.67	.69	-.23	.40	-.02	.19	.16	-.32	.35	.36	.33
2		.55	.35	.45	.53	.18	.41	.17	.44	.49	-.27	.21	.40	.49
3			.59	.61	.16	-.02	.34	.03	.26	.51	-.21	.16	.51	.73
4				.58	.46	.04	.21	.32	.05	.28	-.05	.44	.38	.27
5					.19	.31	.10	.21	-.22	.34	-.25	.48	.48	.53
6						.07	.63	.55	.31	-.06	.26	.38	.32	.13
7							-.01	.41	-.08	.45	.43	.26	.14	-.00
8								.29	.04	-.10	.10	-.08	.48	.24
9									.04	.10	.47	.45	.59	.11
10										.42	.03	.14	.00	.37
11											-.22	.22	.09	.50
12												.26	.17	-.29
13													.67	.29
14														.48

*r = .46 for p .05 and
r = .58 for p .01

Table 6
Performance Specification Criteria (Stabilization) for Performance Tests

TEST		MEANS	STANDARD DEVIATIONS	STABILITY OF CORRELATIONS	OVERALL STABILITY
Stroop	BW	Asymptote	Level	Yes	Yes
	CB	Asymptote	Level	Yes	Yes
	CW	Asymptote	Level	Yes/Marginal	Yes
	BW-CB	Asymptote	Level	No	No
	CB-CW	Asymptote	Level	No	No
Time Estimation		Plateau	Slow Increase or Level	No	No
Complex Counting		Plateau	Level	No/Marginal	No
Critical Tracking		Slow Increase	Level	Yes/Marginal	Yes
Arithmetic		Slow Increase	Slow Increase	Yes	Yes
Air Combat Maneuvering		Slow Increase	Slow Increase or Level	Yes	Yes

FIGURES

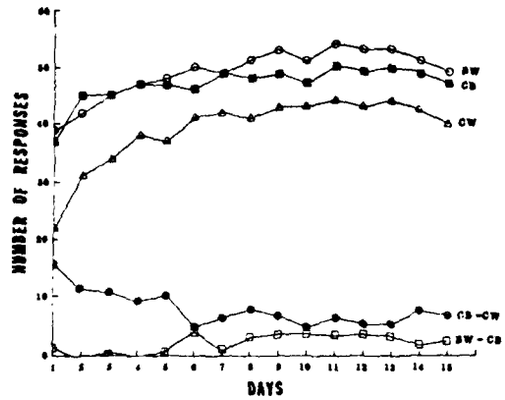


Figure 1. Stroop Test: Group means for black and white words (BW), color blocks (CB), color words (CW), BW-CB, and CB-CW over 15 days (n=19).

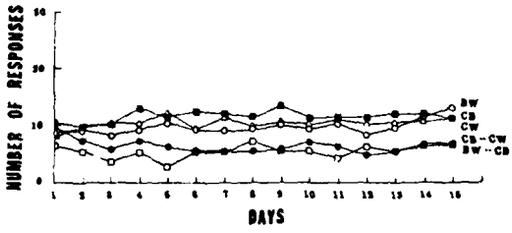


Figure 2. Stroop Test: Standard deviations for BW, CB, CW, BW-CB and CB-CW over over 15 days (n=19).

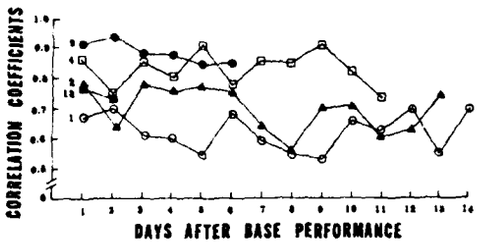


Figure 3. Stroop Test: BW reliabilities for selected base days (1, 2, 4, 9 & 13) and those following over 15 days (n=19).

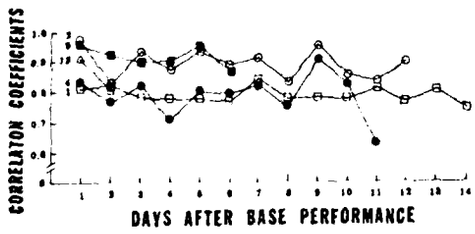


Figure 4. Stroop Test: CB reliabilities for selected base days (1, 2, 4, 9 & 13) and those following over 15 days (n=19).

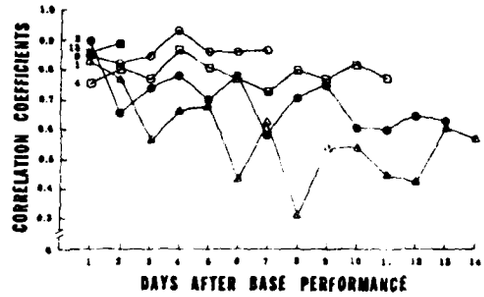


Figure 5. Stroop Test: CW reliabilities for selected base days (1, 2, 4, 9 & 13) and those following over 15 days (n=19).

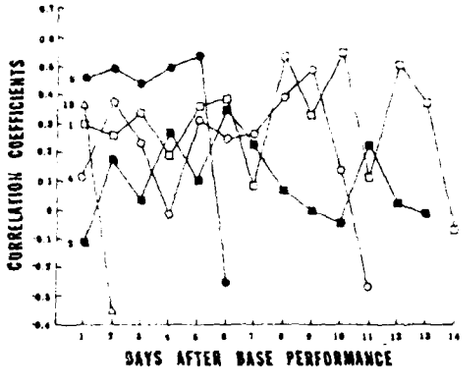


Figure 6. Stroop Test: BW-CB reliabilities for selected base days (1, 2, 4, 9 & 13) and those following over 15 days (n=19).

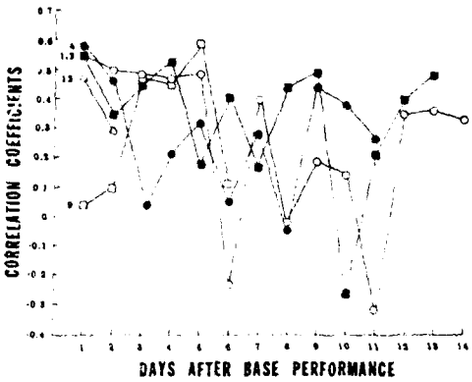


Figure 7. Stroop Test: CB-CW reliabilities for selected base days (1, 2, 4, 9 & 13) and those following over 15 days (n=19).

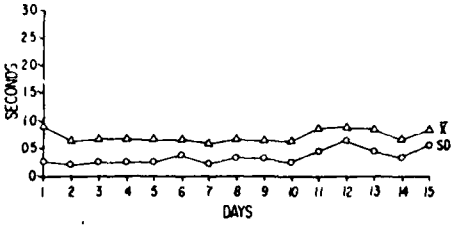


Figure 8. Time Estimation Test: Group means ($\bar{X}$) and standard deviations (SD) for constant error score over 15 days (n=19).

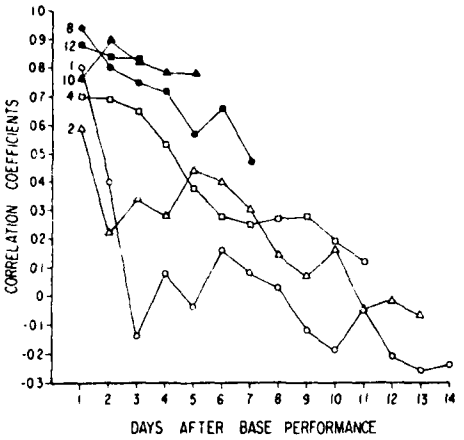


Figure 9. Time Estimation Test: Reliabilities of constant error score for selected base days (1, 2, 4, 8, 10 & 12) and those following over 15 days (n=19).

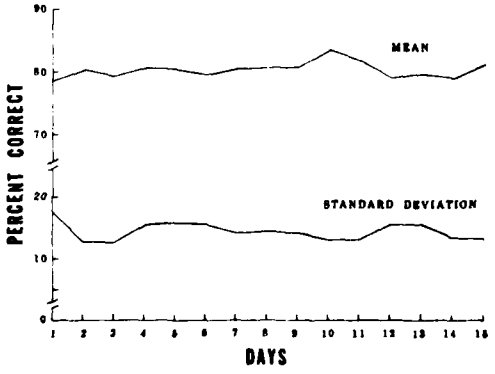


Figure 10. Complex Counting Test: Group means ($\bar{X}$) and standard deviations (SD) for percent correct over 15 days (n=19).

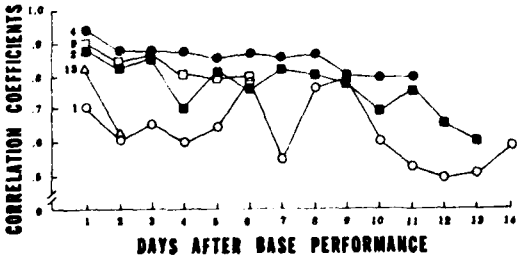


Figure 11. Complex Counting Test: Reliabilities of percent correct for selected base days (1, 2, 4, 9 & 13) and those following over 15 days (n=19).

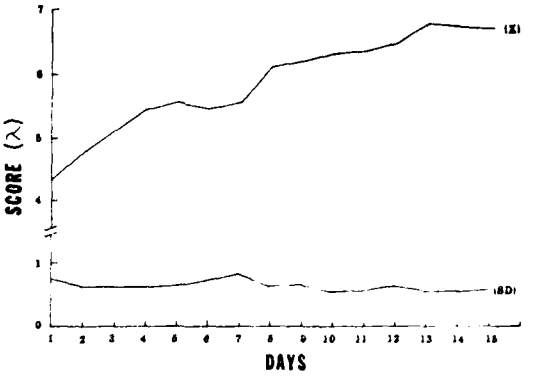


Figure 12. Critical Tracking Test: Group means ($\bar{X}$) and standard deviations (SD) over 15 days (n=18).

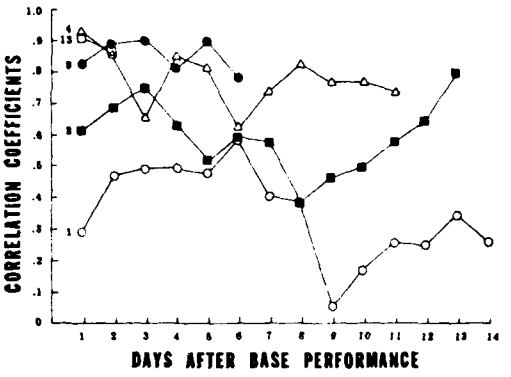


Figure 13. Critical Tracking Test: Reliabilities for selected base days (1, 2, 4, 9 & 13) and those following over 15 days (n=18).

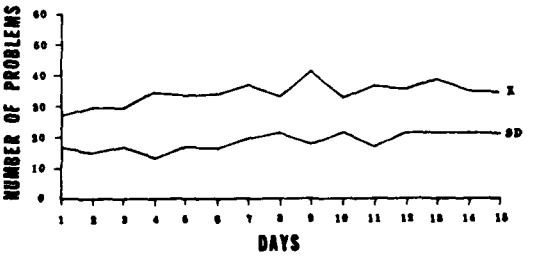


Figure 14. Arithmetic Test: Group means ($\bar{X}$) and standard deviations (SD) for total correct over 15 days (n=18).

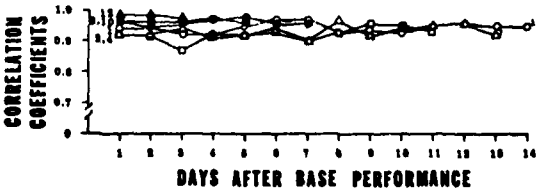


Figure 15. Arithmetic Test: Reliabilities of total correct for selected base days (1, 2, 4, 8, 10 & 12) and those following over 15 days (n=18).

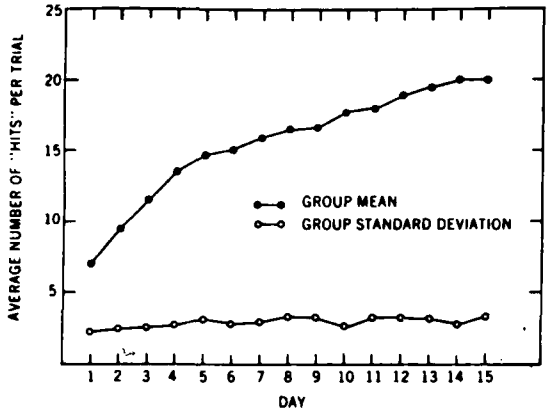


Figure 16. Air Combat Maneuvering Test: Group means and standard deviations for number of hits over 15 days (n=13).

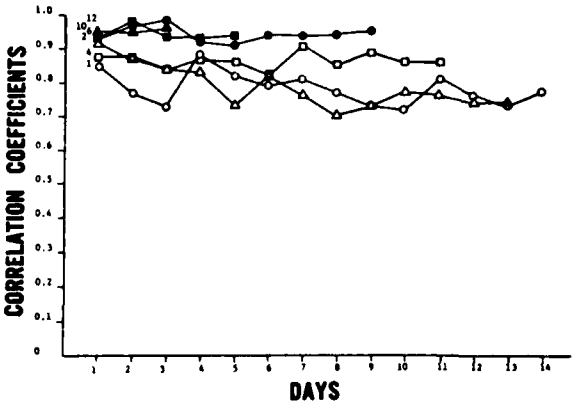


Figure 17. Air Combat Maneuvering Test: Reliabilities of number of hits for selected base days (1, 2, 4, 6, 10 & 12) and those following over 15 days (n=13).

PROCEEDINGS OF THE 24TH ANNUAL MEETING OF THE HUMAN FACTORS SOCIETY
LOS ANGELES, CA, 13-17 OCTOBER 1980

PERFORMANCE EVALUATION TESTS FOR ENVIRONMENTAL RESEARCH (PETER): AUDITORY DIGIT SPAN TASK[1]

Denise B. McCafferty[2]
University of West Florida

Alvah C. Bittner, Jr. and Robert C. Carter
Naval Biodynamics Laboratory
New Orleans, LA 70189

ABSTRACT

Auditory digit span was evaluated as an instrument for repeated measurements experimentation. Twelve subjects were tested for one hour on each of 12 consecutive workdays in a standard environment. Both forward and backward digit span were measured. It was found that forward digit span was suitable for repeated measures after ten days of practice at 30 minutes per day. The criteria for suitability were predictability of the mean scores, constancy of the standard deviations and differential stability of the intertrial correlations. These criteria are sufficient conditions both for repeated measures Analysis of Variance, and for interpretation of experimental effects. Although the backward digit span scores did not meet these criteria, they became more and more correlated with the forward digit span scores as the experiment progressed. This indicates that the mental content of the two tests of memory converged with practice. One implication of this finding is to question the meaningfulness of factor structure after only limited practice. The forward auditory digit span test was recommended for inclusion in a battery of Performance Evaluation Tests for Environmental Research (PETER).

INTRODUCTION

Background

Tests of human cognitive and psychomotor ability are being evaluated for inclusion in a battery of Performance Evaluation Tests for Environmental Research (PETER). PETER is a human performance task battery which is being specifically designed by the Naval Biodynamics Laboratory for repeated administration in unusual environments (e.g., ship motion, vibration, hyperbaria, thermal extremes, drug administration) (Kennedy & Bittner, 1977; Kennedy, Bittner, & Harbeson, 1980). Candidate tests must meet at least one of the following criteria: (1) measure some aspect of information processing; (2) be neurophysiologically diagnostic, or (3) show sensitivity to unusual environments (Kennedy & Bittner, 1977; Kennedy, et al., 1980).

Before tasks are included, they must be found suitably stable for simple analysis and interpretation. Kennedy et al. (1980) and Jones (1980) have suggested that stability exists when: (a) mean performance reaches an asymptote or evidences a slight constant slope, (b) day-to-day variance is constant, and (c) relative performance standings among subjects are constant from day to day, as indicated by unchanging intertrial correlations (differential stability). The first of these stability criteria, for the means, was indicated by Campbell and Stanley (1966) as required for meaningful interpretation of experimental results. The latter two, for variances and correlations, were derived from the sufficient (covariance) matrix condition for repeated measures Analysis of Variance (Winer, 1971). PETER requires all three of these stability criteria.

Purpose

The present study was undertaken to determine whether baseline performance on Auditory Digit Span (ADS) (Ekstrom, French, Harman, and Derman, 1976; Wechsler, 1958) would stabilize following repeated administration of both ADS forward (DF) and backward (DB).

METHOD

Subjects

Subjects were 9 healthy Navy enlisted males (ages 18 to 25) assigned to the Naval Biodynamics Laboratory, New Orleans, as full-time volunteer research subjects. All volunteer research subjects were recruited, evaluated and employed in accordance with procedures specified in Secretary of the Navy Instruction 3900.39 and Bureau of Medicine and Surgery Instruction 3900.6. These instructions are based upon voluntary informed consent, and meet provisions of prevailing national and international guidelines. Each subject was selected for his mental and physical ability to withstand possible hazardous environmental research. Subjects were, however, considered representative of the Navy enlisted population in intelligence (c.f. Thomas, Majewski, Ewing, & Gilbert, 1978).

Apparatus

The Ekstrom et al. (1976) Auditory Number Span Task, based on the seminal work of Kelly (1954), was used as a model for the development of 52 alternate forms, 28 DF and 24 DB. In accordance with Ekstrom et al. (1976), each form consisted of 24 separate series of digits. Each series contained between 4 and 12 digits.

---

1 This research was performed under Navy Work Unit No. MF58.524.002-5027. The opinions are the authors' and do not necessarily reflect those of the Department of the Navy.

2 Now at the Essex Corporation, Alexandria, VA.

The 28 DF and 24 DB tests were randomly assigned to the 12 days of presentation. The four extra forms of DF were used during a two-day pilot experiment which immediately preceded the 12 days of the main experiment. In the following discussion, the third day of exposure to DF testing will be called Day 1 so that the results for DF and DB can be described on a common time line. Each day of testing consisted of 2 different forms of DF and DB, so that the within-session reliability of the tests could be assessed.

Readings of the lists were recorded on an Ampex® 600 reel-to-reel tape recorder using Ampex® 641 magnetic tape. Wechsler's method of reading one digit per second with a drop in voice inflection on the last digit in a series was used (c.f. Hagen, Durham, & Shannon, 1977).

Procedure

Subjects were tested in groups between 0745 and 0845 on 12 consecutive workdays. Prior to the experiment, orientation to the task was held which involved an explanation of the task, instructions, and task demonstrations.

Sessions consisted of four 15 minute sections, two DF and two DB. Instructions were given prior to each section. On the DF portion of the task, subjects were instructed to listen to tape recorded numbers. Upon cue, they were to write those numbers on their answer sheets in the exact order in which they were presented. (Response time of 2 seconds per presented digit was allotted). Following Ekstrom et al. (1976), a subject's scores were the number of correctly recorded series. Therefore, scores for DF or DB could range between 0 and 48 for the two forms composed of 24 series each. Instructions for DB were the same, with the exception that subjects were instructed to write their answers in the exact opposite order to which they were called out.

RESULTS

The data were analyzed in two phases. During the first phase, the DF and DB tasks were checked for stability. The structure of the forward and backward portions of the test were compared during the second phase.

Task Stability

Means and Standard Deviations. Figure 1 shows the average DF and DB scores over days. It appears that DF means are larger than DB and that the difference is constant. Table 1 supports this view with a significant difference between the means for total forward and total backwards and with no interaction of DF versus DB and days. The effect of days, it is noteworthy, also was significant reflecting a gain in performance over trials which is approximately linear after Day 4 (for nonlinear trends, $F(7,21) = 2.17$, $p > .08$). The slope of the linear trend is 0.11 series per day.

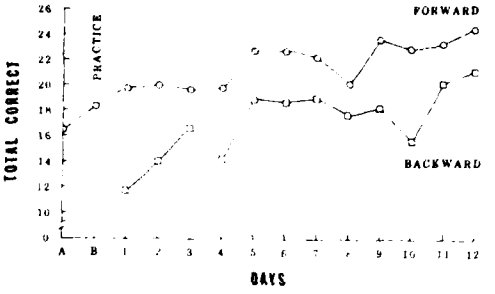


Figure 1. Mean Total Correct Forward and Backward Digit Span for 12 Days (N=9)

TABLE 1

ANALYSIS OF VARIANCE (ANOVA)

	DF	SS	F	P
DF vs DB INSTRUCTION (I)	1	1148.44	88.00	10^{-10}
DAYS	11	931.09	5.49	10^{-8}
I X D	11	162.30	1.13	n.s.
SUBJECTS	8	4177.29	40.02	10^{-10}
RESIDUAL	184	2400.49		
TOTAL	215	8819.70		

An Fmax test comparing the largest to the smallest within-day variances on the DF and DB tasks found no significant difference for forward and backward tasks respectively, ($Fmax_F = 6.25$, and $Fmax_B = 5.92$, $p > .10$).

Intertrial Correlations. Figures 2 and 3 demonstrate the pattern of the DF and DB correlations between scores obtained on days near the onset of testing and those obtained on later days. These figures show correlations of scores

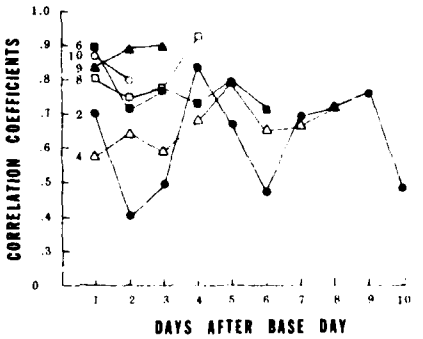


Figure 2. Correlations Between Selected Base Days and Following Days for Total Correct Forward Digit Span for 12 Days (N=9)

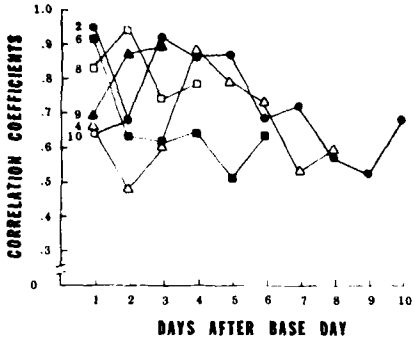


Figure 3. Correlations Between Selected Base Days and Following Days for Total Correct Backward Digit Span for 12 Days (N=9)

obtained on selected Base Days with scores obtained 1, 2, or more days later. This method is helpful in demonstrating not only the reliability of the task over time, but also, in the case of DF, differential stability. If these plots were flat and overlapping after some day in practice, the test scores were considered to be stable after that day (Bittner, 1979). This pattern is suggested by the lines representing Base Day 9 and following days on the "total forward" graph. Conversely, the downward slope of the lines on the "total backward" graph does not indicate stability. Lawley tests (Morrison, 1963) supported the view of the graphical analyses. In the case of backwards ADS, the Lawley test indicated significant instability of the intertrial correlation matrix ($\chi^2 = 8.07$, df = 2, $p < .01$) across even the last 3 days of the study, after 9 days of practice. In contrast, stability was found for forward ADS on the 4 days after Day 8 ($\chi^2 = 8.377$, df = 5, $p < .10$). If the two extra days practice for DF are considered, we can conclude that DF stabilizes after 10 days of practice for ½ hour per day. The intertrial correlation matrices for DF and DB are presented in Tables 2 and 3.

TABLE 2

AUDITORY DIGIT SPAN TASK: Inter-day Correlations for Forward Task Over 12 Days (N=9)

DAYS	2	3	4	5	6	7	8	9	10	11	12
1	58*	83	58	79	84	93	60	53	68	60	60
2		70	40	49	84	67	47	69	72	76	48
3			71	54	82	84	51	57	59	66	49
4				57	64	59	68	79	65	67	72
5					79	89	84	75	75	79	91
6						89	72	77	73	79	71
7							74	66	72	79	76
8								80	75	78	93
9									83	90	89
10										87	80
11											88

* Decimal Points Omitted

TABLE 3

AUDITORY DIGIT SPAN TASK: Inter-day Correlations for Backwards Task Over 12 Days (N=9)

DAYS	2	3	4	5	6	7	8	9	10	11	12
1	68*	67	63	71	44	32	35	35	19	-04	21
2		95	68	92	86	87	68	70	57	52	68
3			64	94	89	86	64	57	55	39	54
4				67	48	60	88	79	73	53	59
5					83	84	63	67	48	48	63
6						92	62	61	64	52	63
7							76	78	72	72	77
8								83	94	74	79
9									69	87	89
10										64	68
11											95

* Decimal Points Omitted

Task Structure

The second portion of analysis was devoted to the examination of the structure of the forward and backward tasks by graphical analysis and analysis-of-variance. Reliabilities for each day's total scores were obtained for each task. For each day, the square root of the product of the two tasks' reliabilities was then plotted on a graph to represent the maximum expected correlation between DF and DB (see Figure 4). The maximum theoretical correlations, it is noteworthy, would be obtained when all of the reliable variance on both DF and DB tasks measures a single "*factor*" (*Harman*, 1975). *The correlations* of digit span forward and backward were also plotted on this graph. This was an attempt to show the relationship between the tasks, given the maximum possible correlation allowed by the reliabilities. It is clear that the correlation between the ADS tasks approaches the maximum possible as trials progress. Hence, DF and DB converge with practice.

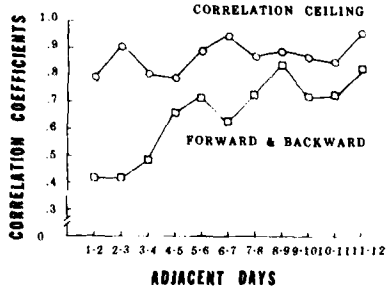


Figure 4. Forward and Backward Digit Span Correlations across Adjacent Days Compared with Correlation Ceilings (N=9)

In contrast to the convergence of content illustrated by the correlation results, Figure 1 demonstrates the unchanging difference of difficulties across days represented by the means of forward and backwards. As noted before, there was no significant interaction of instruction (forward-backwards) and day. This indicates that the effects of instruction and experience with ADS are additive and independent. In summation, the forward and backward ADS processes appear to become more similar in content with practice but their means remain different by a constant amount.

DISCUSSION

Task Suitability for Performance Tests

The forward portion of the auditory digit span task (DF) was found to be suitable for the PETER battery. In particular, the change of the DF mean performance was found to be approximately linear after Day 4 and the variances evidenced no significant change over the course of the study. In addition, the DF task was found to be differentially stable for the last four days of testing. The reliability of the DF task, it is pertinent to note, was comparable to that ($\underline{r}$ = .74) reported by Ekstrom, et al. (1976) with $\underline{r}$ = 0.86 over the differentially stable days. The DF task meets the suitability criteria for means, variances, and correlations required for simple analysis and interpretation (Kennedy & Bittner, 1977; Kennedy, et al., 1980).

In contrast to DF, the backward task (DB) failed to stabilize suitably for consideration for inclusion in PETER. While the analysis of means showed a linear increase after Day 4 and constant variances, the task did not evidence differential stability even after 9 days of practice. The average reliability of DB for the last three days was moderately high with $\underline{r}$ = 0.76, suggesting ultimate reliability in the neighborhood of that seen for the DF task. However, convergence of the DB on the DF task, as seen in Figure 4, suggests that the DB task would eventually become stable as it continued to approach the DF task. This approach to differential stability is also suggested by the slopes of the traces seen in the graphical analysis of Figure 3. The slopes of the traces appear to be approaching a zero slope as base days become later. The ultimate convergence of the DB task to differential stability is an empirical question which requires more investigation. The DB task is unlikely, however, to be of interest for a task battery as DB appears to be approaching DF which is already stable.

Implications for Performance Testing

The implications of the results for performance testing are twofold. First, the stability of the mean and standard deviations after the fourth day would have misled investigators who use only these statistics for determining the suitability of a task before beginning an environmental investigation. The changing character of what-is-being-measured (Alvares & Hulin, 1972), as indicated by the intertrial correlations, would not have been apparent to such investigators and the meaningfulness of their results would have been unknowingly compromised (Bittner, 1979). This would be particularly true when the magnitude of change of the intertrial correlations is as large as the changes reported by other investigators (Kennedy, et al., 1980). The second implication of this investigation's results is to question the meaningfulness of task differences seen with only one or two trials of practice. In early stages of training, DF and DB tasks measured non-overlapping variance. However, with more training, the overlap was seen to increase. How true this would be for other tasks currently believed to measure distinct abilities is an empirical question. Current factor batteries (e.g., Ekstrom, et al., 1976) have been developed based on performances with no or only one trial of practice. The possibility is suggested that, with repeated testing, the plethora of human performance factors or abilities may converge to far fewer than presently thought. Both implications for performance testing revolve around the issues of changes in the character of a task with practice. The issues deserve greater attention and investigation.

Conclusion

The forward portion of the auditory digit span task is suitable for use in environmental research employing repeated measures. Auditory Digit Span is recommended for inclusion in a test battery as a measure of inattention or freedom from distraction and as an indicator of short term memory or neurophysiological impairment.

REFERENCES

Alvares, K. M., & Hulin, C. L. Two explanations of temporal changes in ability-skill relationships: A literature review and theoretical analysis. Human Factors, 1972, 14, 295-308.

Bittner, Jr., A. C. Statistical tests for differential stability. Proceedings of the 23rd Annual Meeting of the Human Factors Society. Boston, October 1979.

Campbell, D. T. & Stanley, J. C. Experimental and quasi-experimental designs for research. Chicago: Rand McNally, 1966.

Ekstrom, R. B., French, J. W., Harman, H. H., & Derman, D. Manual for kit of factor referenced cognitive tests. Princeton, N.J.: Educational Testing Service, 1976.

Hagen, R. L., Durham, T., & Shannon, D. Administration of digit span on Wechsler and Binet: Differences. Journal of Clinical Psychology, 1977, 33, 480-482.

Harman, H. H., Modern factor analysis. Chicago: University of Chicago Press, 1975.

Jones, M. B. Stabilization and task definition in a performance test battery. (NBDL Monograph No. M-0001) New Orleans, LA: Naval Biodynamics Laboratory, 1980.

Kelley, H. P. A factor analysis of memory ability (Ph.D. thesis, Princeton University, 1954). Educational Testing Service Research Bulletin, 7, 1954.

Kennedy, R. S., & Bittner, Jr., A. C. The development of a performance evaluation test for environmental research (PETER). In L. T. Pope & D. Meister (Eds.), Productivity Enhancement: Personnel Performance Assessment in Navy systems. Symposium presented at the Naval Research and Development Center, San Diego, October 1977, 393-408.

Kennedy, R. S., Bittner, Jr., A. C. & Harbeson, M. M. An engineering approach to the standardization of performance evaluation tests for environmental research (PETER). Proceedings of the 11th Annual Conference of the Environmental Design Research Association, Charleston, SC, 2-6 March, 1980.

Morrison, D. F. Multivariate statistical methods. New York: McGraw-Hill, 1967.

Thomas, D. J., Majewski, P. L., Ewing, C. L., & Gilbert, N. S. Medical qualification procedures for hazardous duty. Aeromedical Research Conference Proceedings (No. 231 A3). London: AGARD, 1978, 1-13.

Wechsler, D. Measurement and appraisal of adult intelligence, (4th ed). Baltimore: Williams & Wilkins, 1958.

Winer, B. J. Statistical principles in experimental design. New York: McGraw-Hill, 1971.

PROCEEDINGS OF THE 24TH ANNUAL MEETING OF THE HUMAN FACTORS SOCIETY
LOS ANGELES, CA, 13-17 OCTOBER 1980

COMPARISON OF MEMORY TESTS FOR ENVIRONMENTAL RESEARCH

Mary M. Harbeson, Michele Krause, and Robert S. Kennedy
Naval Biodynamics Laboratory, New Orleans, LA 70189

ABSTRACT

Four memory tests were considered for inclusion in a human performance test battery. The tests were administered to 23 Navy enlisted men for 15 consecutive days. Group means, standard deviations, and cross-session correlations were examined. Two of the tests, Interference Susceptibility and Free Recall, met the initial statistical criteria for inclusion in the test battery. However, the other tests, Running Recognition and List Differentiation failed to show sufficient task definition and reliability in their present form. These tests are compared with each other and with previous memory research studies.

INTRODUCTION

Memory functions are among the complex mental operations which are involved in Navy jobs and play a role in the effectiveness of Navy systems. This report focuses on an evaluation of four memory tests which were considered for the Performance Evaluation Tests for Environmental Research battery. Comparisons are made between the present study and research by Underwood, Boruch, and Malmi (1977) and Fernandes and Rose (1978) in which the same tests were examined for different purposes. Present efforts are devoted to the development of a test battery which will be used to determine the extent of performance decrements in stressful environments (Harbeson, Kennedy, & Bittner, 1979; Kennedy & Bittner, 1977, 1978; Kennedy, Bittner, & Harbeson, 1980; Kennedy, Carter, & Bittner, 1980). Cognitive, perceptual, and psychomotor tests which were previously shown to be sensitive to several validity criteria have been selected for study (Carter, Kennedy, & Bittner, 1980; Kennedy, Carter, & Bittner, 1980). Tests meeting these initial criteria are administered and evaluated to determine whether they are stable and reliable after extended practice. Future research will employ real world work criteria from task analyses (Shannon, 1980a) in order to select and validate subsequent tasks.

The strategy of this research program has been to administer each task for 15 consecutive workdays to the same group of subjects. Tasks are considered stable if after practice: (a) the means are level or evidence a slight, constant slope over days, (b) the standard deviations are level, and (c) the between trial correlations cease to change over trials. In addition, cross-session reliabilities (task definition) must be high enough to differentiate among individuals. A correlation of .707 has been set as the lower limit for acceptability. Tests which are both stable and have adequate task definition are selected for tentative inclusion in the test battery.

The four memory tests in this study were adapted from Fernandes and Rose and based on the earlier work of Underwood, Boruch, and Malmi. Each test was designed to measure a different aspect of memory. Free Recall was designed to measure recall or retrieval skill. Running Recognition dealt with recognition or the ability to discriminate between memories. List Differentiation was used as a measure of temporal discrimination, and Interference Susceptibility was designed to study the effects of proactive interference. These tasks were selected as representative of a larger body of tasks studied by Underwood, et al.

The authors examined the interrelationships among 28 episodic and 5 semantic memory tasks in order to determine the correlations among various attributes of memory (associative, temporal, acoustic, etc.). Each task was administered once, to 200 college students. A factor analysis revealed 5 factors: (a) paired-associate/serial, (b) free recall, (c) memory span, (d) recognition/frequency discrimination, and (e) verbal discrimination. These factors were related to the tasks rather than to the attributes.

Fernandes and Rose selected 6 tests from the Underwood, et al. study. These authors were interested in an information-processing approach to the problems of both individual differences and memory function. Their objective was an assessment instrument that could be generalized to a wide range of criterion tasks. Each test was administered twice, to 22 office workers. Fernandes and Rose employed the Underwood stimulus material for their first session, and generated equivalent alternate forms for the second session. The results of their study led Fernandes and Rose to propose 5 of the 6 tests as candidates for their performance battery, omitting Interference Susceptibility because of extreme variations in group performance. They further commented that the memory tests appeared more related to general skill in encoding and storage than to the attributes they were nominally purporting to measure.

In the present study four of the six tests used by Fernandes and Rose were administered for 15 consecutive working days. Situational Frequency was excluded because it did not lend itself to easy construction of alternate forms. Because of time constraints, Digit Span, a task similar to the Memory Span Test suggested by Fernandes and Rose, was administered to a different population and is reported separately (McCafferty, Bittner, & Carter, 1980).

Purpose

The purpose of this study was to determine the effects of extended practice on four memory tests and to determine their suitability for

inclusion in a human performance test battery.

METHOD

Task Descriptions

Running Recognition. Subjects were shown a long list of words and were asked to indicate whether each item was old or new by circling the appropriate response on their answer sheets. An example of the stimulus presentation is shown in Table 1. This test was based on a test developed by Shepard and Teghtsoonian (1961), who used numbers rather than words. The Underwood group designed their test to measure recognition sensitivity and an acoustic attribute. The test included words of different acoustic characteristics, which were repeated at different lags within a list. Two lists were used, one containing 173 words, and the other 174. Each word was displayed for 4 seconds. Fernandes and Rose used the Underwood, et al. stimuli to construct a list of 101 words for each of their two testing sessions. All words, except one, appeared twice in a list, and lags between the words varied from 1 to 36 words. Each word was displayed for 3 seconds.

TABLE 1

Running Recognition: Example of Stimulus Presentation

STIMULUS	RESPONSE SHEET	
INCOME	(NEW)	OLD
BUILD	(NEW)	OLD
INCOME	NEW	(OLD)
-	-	-
CHATTER	(NEW)	OLD

In the present experiment, the Fernandes and Rose procedure was followed but the lists for each day were reduced to 51 words. Alternate forms were generated by selecting words in a pseudo-random manner from the pool of 101 original stimulus words. There were 5 unique orders of presentation.

TABLE 2

List Differentiation: Example of Stimulus Presentation

STIMULUS MATERIAL			RESPONSE SHEET			
LIST 1	LIST 2	LIST 3				
prow	swab	soon	need	1	(2)	3
cost	meet	area	thaw	1	2	(3)
miss	adds	thaw	cost	(1)	2	3
-	-	-	-	-	-	-
foil	that	atop	area	1	2	(3)

List Differentiation. Three distinct lists of four-letter words were presented. The same words were arranged in random order on the response sheets, followed by the digits 1, 2, and 3. The subjects were required to indicate the list to which each item belonged (see Table 2). Underwood, et al. administered 2 sets of 3 lists each, with 20 words per list in 1 session. The response time was unpaced. Fernandes and Rose and the present study followed this procedure except that the response time was set at 3 minutes, and in the present study only 1 set of 3 lists per day was used.

Free Recall. The Fernandes and Rose stimulus material was used and additional alternate forms were generated following the Underwood, et al. method. Subjects were shown lists of common words and were instructed to write as many as they could remember on their answer sheets. Three conditions were used: control, concrete, and abstract. An example of the stimulus material is shown in Table 3. The control condition, which, was described by Fernandes and Rose as a measure of short-term memory, consisted of five-letter words selected at random from the Thorndike-Lorge (1944) tables.

TABLE 3

Free Recall: Example of Stimulus Material

CONTROL	CONCRETE	ABSTRACT
sugar	body	trouble
yield	circle	hour
horse	gentleman	method
-	-	-
quote	arrow	affection

The concrete and abstract conditions were designed to measure encoding by imagery. Words with values above 6 on the Paivio, Yuille, and Madigan (1968) rating scale were used in the concrete condition, and those with values below 3 were used in the abstract condition. Underwood, et al. used 4 lists for the control condition and 2 lists each for concrete and abstract conditions, with 24 words per list. Subjects were shown each word for 4 seconds, with 2 minutes allowed for recall at the end of each list. Fernandes and Rose followed the same procedure as Underwood, et al. except that lists of 20 words were used, and the presentation time and response time were reduced by 50%. In the present study, only 2 lists of control, and 1 list each of concrete and abstract words were used each day. Approximately 30% of the words were used twice, with the contingency that the same word was not repeated on adjacent days. Testing time occupied 7 minutes per session.

Interference Susceptibility. Stimulus material for each session was comprised of paired-associate lists. A list was made up of 5 three-letter words

paired with the digits 1 - 5. Table 4 gives an example of the stimulus presentation. Each set consisted of 4 lists, in which the same words and digits were used, but paired differently and presented in a different order in each list. Five new words were paired with the digits 1 - 5 in each set. After each paired-associate list was presented, the words alone were shown in random order and the subjects were required to write the appropriate digit on their response sheets. Six sets of stimuli were presented in both the Underwood, et al. and Fernandes and Rose studies. Inspection and response times for each item were 3 seconds. In the present study, only 3 sets per session were presented.

TABLE 4

Example of Interference Susceptibility

INSPECTION LIST	TEST LIST	CORRECT RESPONSE
DOG-5	WIN	1
NOB-2	PEG	4
WIN-1	DOG	5
PEG-4	NOB	2
HEW-3	HEW	3

Subjects

The subjects were a group of 23 volunteer enlisted Navy men, ages 19 to 24. To qualify for this medical research program, they had to be average or above the norms for Navy enlisted personnel in physical health, mental health and intelligence. All subjects were recruited, evaluated and employed in accordance with procedures specified in Secretary of the Navy Instruction 3900.39 and Bureau of Medicine and Surgery Instruction 3900.6. These instructions are based upon voluntary consent, and meet the provisions of prevailing national and international guidelines. A description of the subject selection procedure is given by Thomas, Majewski, Ewing, and Gilbert (1978).

Apparatus

The stimulus material consisted of 2 X 2 inch black and white slides with one item per slide presented on a Kodak Ektagraph 450 Audio Viewer®. The rate of presentation was controlled by preprogrammed tape cassettes. Subjects recorded their answers on response sheets.

Procedure

The subjects were tested in groups of four beginning at 8:00 AM for 15 consecutive workdays. The four tests were administered in the same order to each group of subjects, but the order varied for different groups. There was a break of 2 or 3 minutes between tests while the experimenter changed slide carousels and cassette tapes. T. ing lasted approximately 40 minutes per day.

RESULTS AND DISCUSSION

Running Recognition

An overall percent correct score was calculated. Figure 1 shows means and standard deviations. Group means begin at 95% and decrease slightly over days, ranging between 95% and 89%. The average standard deviation is 5.87%, and although variable, did not show any positive or negative trend.

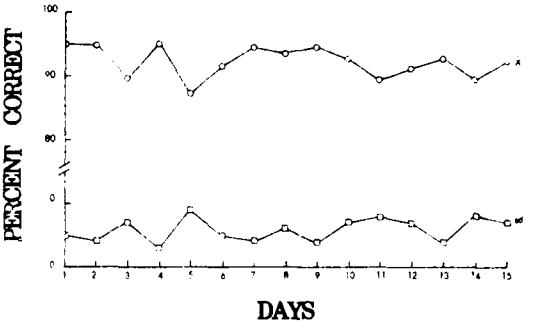


Figure 1. Running Recognition means and standard deviations for percent correct across 15 days (n=23).

The graph of the cross-session correlations which is shown in Figure 2 was constructed by plotting the correlations between a base day and each subsequent day (e.g. Day 1 to 2, 1 to 3, ... 1 to 15). Correlations are extremely variable, but there is no obvious trend. Because task definition is very low ($\bar{r} < .20$), this test does not meet the minimum criteria for inclusion in the performance test battery. It is believed that in the present

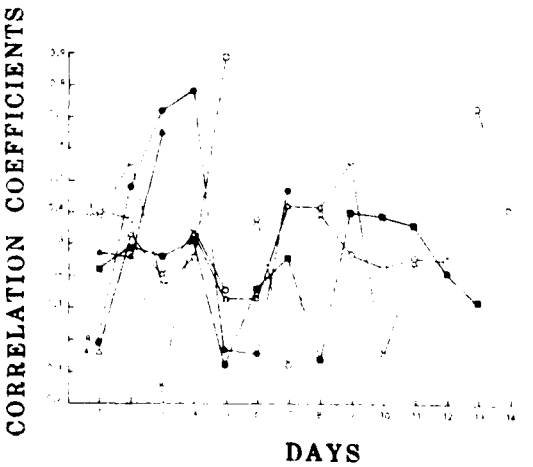


Figure 2. Running Recognition correlation traces for percent correct across 15 days (n=23).

study, shortening the test made it too easy. This may have caused a ceiling effect, which lowered between-subject variance and therefore, reliability. A Spearman adjustment for test length indicated that making our test comparable in length would raise the correlations to what Underwood obtained. However, a 23 minute memory test would be prohibitively time consuming as part of a battery. Possibly, a selection of different stimulus material (e.g., nonsense syllables or abbreviations) would provide the required reliability (sensitivity) with more modest testing time.

List Differentiation

A percent correct score for each of the three lists was calculated. Means and standard deviations for the three lists were comparable. The most reliable score, however, proved to be percent correct across all lists, and this was used in subsequent analyses. Means and standard deviations appear level across sessions (Figure 3). Analysis of Variance and Fmax tests were non-significant.

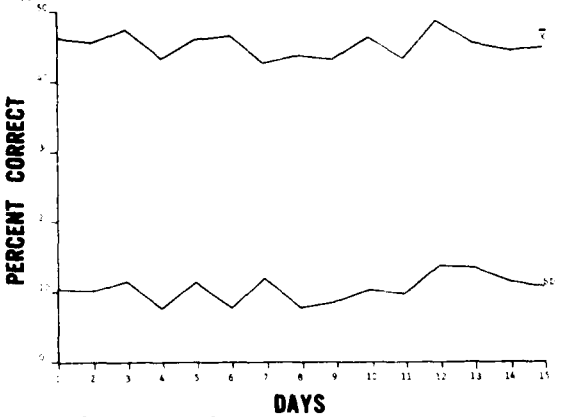


Figure 3. List Differentiation means and standard deviations for percent correct across 15 days (n=23).

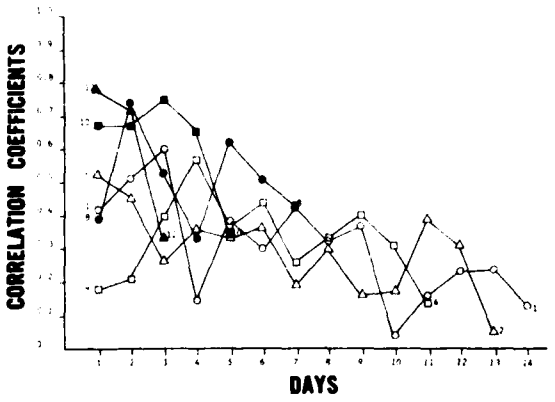


Figure 4. List Differentiation correlation traces for percent correct across 15 days (n=23).

Correlation traces (Figure 4) are generally low ($\bar{r}$ = .37), but improve somewhat with later days. Early traces tend to decline as performance becomes more remote from the base day, reflecting instability. However, with the exception of the final day which was extremely low, (c.f. Shannon, 1980b) there was a tendency for later base days to have higher correlations than earlier days (for Days 9 - 14, $\bar{r}$ = .64). Therefore, this task stabilizes when Day 15 is dropped. In the present study, the shorter testing time (50% of Fernandes' & Rose) and task difficulty may have contributed to the lower correlations (note, the average percent correct score across 15 days was 45.25%). This task is not suitable in its present form, but with modifications (e.g., stimulus material with more meaningful associations), it could be made acceptable for the performance test battery.

Free Recall

Percent correct scores were calculated for the control, concrete and abstract conditions. The means and standard deviations for all conditions followed a similar pattern and as expected, performance was generally best for concrete words and poorest for abstract words. The average score across all conditions was used in the analyses because it was the most reliable and was highly correlated each day with all other scores. The means and standard deviations are shown in Figure 5. With the exception of the first and last days, the means appear level with a gradual increase across sessions. The average percent correct score across days is 35%. A significant days effect is shown in the analysis of variance F (14, 308) = 2.54, $p < .01$. Examination of the orthogonal components revealed a significant quartic (4th order) effect. First and last day, and weekend effects may offer an explanation. The standard deviations appear level across days with a slight increase proportional to the means. An Fmax test showed no statistically significant difference across days.

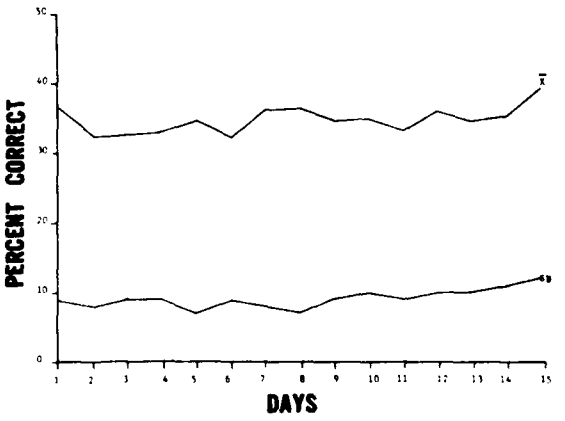


Figure 5. Free Recall means and standard deviations for percent correct across 15 days (n=23).

The correlations for selected base days and those subsequent appear in Figure 6 and reflect no dramatic trend, although there is a tendency for *later day correlations* to be higher than those for earlier days. It appears that the correlations may be stable as early as Day 1. _Task definition', when averaged across 15 days, is $\bar{r}$ = .63 but reaches $\bar{r}$ = .72 when only the days after Day 9 are considered. This task is acceptable for inclusion in the human performance test battery.

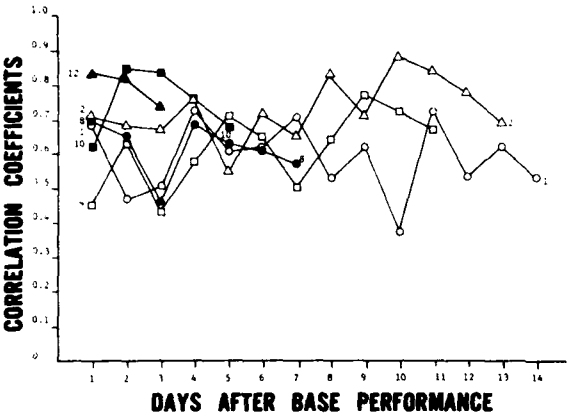


Figure 6. Free Recall correlation traces for percent correct across 15 days (n=23).

Interference Susceptibility

Percent correct scores within each list, across lists, and within and between sets were calculated. In addition, slope scores were calculated across lists. A composite mean score was used, again, because it was the most reliable and because daily part scores correlated highly (generally, $r > .60$) with each other and with the total score. The slope scores, traditional interference measure possessed zero reliability. Figure 7 shows the means and standard deviations. Except for the extremely low score on Day 6, the means show a smooth learning curve which asymptotes after Day 7. The grand mean percent correct is 65%, increasing from a low of 50% on Day 1 to a high of 74% on Day 13. Analysis of variance shows a significant days effect for Days 1 - 15, F (14, 308) = 7.40, $p < .01$, and also for Days 7 - 15, F (8, 176) = 3.13, $p < .01$. This could be explained by the continued and regular increase in performance across sessions. The standard deviations appeared level throughout testing. A non-significant Fmax confirmed this observation.

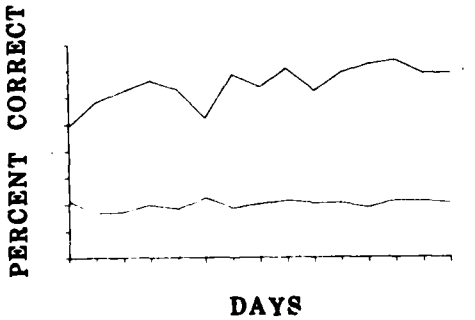


Figure 7. Interference Susceptibility means and standard deviations for percent correct across 15 days (n=23).

The correlation traces (Figure 8) appear to follow a pattern which is to be expected when performances improve with practice. Like the means, Day 6 correlations are anomalous and while the cause is unclear, most probably reflect procedural or apparatus problems. With this exception, the traces appear to be fairly level for each day with the days which follow and increase in value for subsequent base days. The figure has a layered appearance with traces for later days being approximately parallel, and higher than those for earlier days. For the days after Day 7, the traces appear to overlap, indicating stability. The average correlation for Days 7 - 15 is .73, as opposed to .46 overall. This test appears acceptable for use in a human performance test battery. It should be noted, however, that since the measure of interference (slope) had a zero reliability, the specific memory attribute being measured by this test is in question.

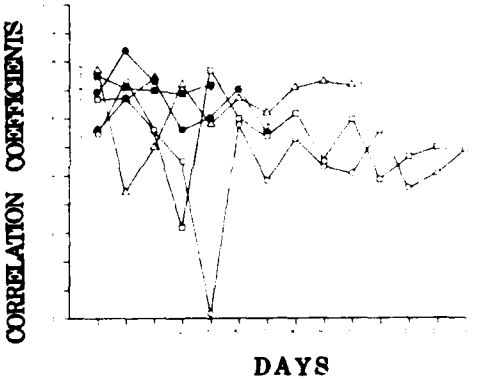


Figure 8. Interference Susceptibility correlation traces for percent correct across 15 days (n=23).

Comparison of Tests

A comparison of results from the present study and past research on these tests is shown in Table 5. Data from the past studies shown in this table were approximated from the published results. In cases where no reliabilities were given for a total score, the reliabilities for each condition were averaged.

For the most part, means and standard deviations in the present study are comparable but tend to be lower than those previously obtained. Running Recognition (RR) has significantly lower correlations. Correlations for List Differentiation (LD) are low when only Days 1 and 2 are examined. However, when days after stability are considered, correlations approach those in past studies. Interference Susceptibility (IS) in the present study reveals higher means for stable days than

those obtained by Fernandes and Rose but lower than those obtained by Underwood et al. In the case of Free Recall (FR), means are substantially lower than those in the Underwood, et al. study, but are essentially the same as those in the Fernandes and Rose study. Different presentation times, 4 seconds in the Underwood, et al. study and 2 seconds in the other two studies, may account for the discrepancy. In general, the differences between this study and past research may be attributed to (a) decreased test length, (b) modifications in the testing procedure, (c) repetition of stimulus material, and (d) subject population differences. The sample used in the present study is representative of the Navy enlisted population. In addition, they are comparable to the general population on at least one measure, the Wonderlic Personnel Test. Even so, it is expected that the college student population in the Underwood, et al. study may be brighter and would be more practiced at tests involving verbal ability. The lower reliabilities that we obtained are probably the consequence of attempting to shorten the tests so that they could all be accomplished within a daily session lasting approximately 30 minutes. It is our opinion that the selection of more relevant (e.g., job related) but more difficult (e.g., abbreviations/acronyms) material may permit shorter tests at no sacrifice to reliability. This will be attempted in a future study.

TABLE 5

Comparison of Three Studies

	Underwood et al. (n = 200)	Fernandes & Rose (n = 22)	Present Study (n = 23)	
Sessions	1	1&2	1&2 (Stable Days)	
RR				(1-15)
Test Time*	23	5	2½	
$\bar{X}$ (%)	93	93	94	91
r (x 100)	70	82	30	18
LD				(10-14)
Test Time	7	7	4	
$\bar{X}$ (%)	55	50	46	46
r (x 100)	71	77	42	64
FR**				(9-15)
Test Time	29	13	7	
$\bar{X}$ (%)	53	38	34	35
r (x 100)	67	77	68	72
IS				(7-15)
Test Time	12	13	6	
$\bar{X}$ (%)	85	65	54	70
r (x 100)	81	77	60	73

N.B. Caution should be taken in interpreting results from tests of different lengths.
* Minutes
** Underwood, et al. used lists of 24 words, whereas the other studies used 20 words per list.

In Table 6, the correlations which appear in the diagonal are the composite of stabilized days within a test. Similarly, the between test correlations which appear in the other cells are also only for stabilized days. Thus, reliability correlations for List Differentation are the arithmetic average of 10 comparisons (Days 10-14) and Free Recall, 21 comparisons (Days 9-15). Moreover, the composite correlations between these two tests are the average of 35 comparisons (i.e. days 10-14 versus 9-15).

TABLE 6

Intercorrelation of Stable Periods of Four Memory Tests

	IS	FR	LD	RR
IS	.73	.50	.32	.25
FR		.72	.51	.17
LD			.64	.21
RR				.18

An inspection of this table reveals correlations between stabilized trials that are higher than the factor analysis of Underwood, et al. would predict since the tests were originally selected for orthogonality. Indeed, given the average low retest reliability of Running Recognition, the present matrix implies only a single factor for all four tests. When calculations were performed over earlier (unstabilized) trials the data were more in line with the low correlations between tests found by Underwood, et al. However, when Days 7-14 of three of the tests (List Differentiation, Free Recall, and Interference Susceptibility) were factor analyzed by Shannon (1980a) 63 percent of the common variance was explained by one factor. These data suggest that following extended practice on a family of tests, a general factor which underlies all the tests may appear. We have had this experience previously in our laboratory (McCafferty, et al, 1980; Kennedy, Bittner, & Jones, 1980). The practical consequences of outcomes like this imply that samples of practiced behavior may have far broader generalizability than was previously thought.

CONCLUSIONS

In conclusion, of the four tasks considered for inclusion in a human performance test battery, Interference Susceptibility and Free Recall were found to be acceptable. List Differentiation and Running Recognition were not acceptable in their present forms but could possibly be useful if modified. The performance on the four tasks was generally comparable, but poorer than that obtained in the previous studies. In addition, it is suggested that with extended practice all four tasks may measure a single factor.

REFERENCES

Carter, R. C., Kennedy, R. S., & Bittner, A. C., Jr. Selection of Performance Evaluation Tests for Environmental Research. Proceedings of the 24th Annual Meeting of the Human Factors Society, Los Angeles, October 1980.

Fernandes, K. & Rose, A. M. An Information Processing Approach to Performance Assessment: An Investigation of Encoding and Retrieval Processes in Memory. (Tech. Report IAR 58500-11/78 TR). Washington, D.C.: American Institutes for Research, November, 1978.

Harbeson, M. M., Kennedy, R. S., & Bittner, A. C., Jr. A comparison of the Stroop Test to other tests for studies of environmental stress. Proceedings of the 12th Annual Meeting of the Human Factors Association of Canada, Bracebridge, Ontario, Canada, September, 1979, 21.1, 21.9

Kennedy, R. S., & Bittner, A. C., Jr. The development of a Navy Performance Evaluation Test for Environmental Research (PETER). In Pope, L. T. & D. Meister, (Eds.), Productivity Enhancement: Personnel Performance Assessment in Navy Systems. Symposium presented at the Naval Personnel Research and Development Center, San Diego, CA, October 1977, 393-408. (NTIS No. AD 056047).

Kennedy, R. S., & Bittner, A. C., Jr. Progress in the analysis of Performance Evaluation Tests for Environmental Research (PETER). Proceedings of the 22nd Annual Meeting of the Human Factors Society, Detroit, Michigan, October, 1978. (NTIS No. AD A060676).

Kennedy, R. S., Carter, R. C., & Bittner, A. C., Jr. A catalogue of Performance Evaluation Tests for Environmental Research. Proceedings of the 24th Annual Meeting of the Human Factors Society, Los Angeles, October, 1980.

Kennedy, R. S., Bittner, A. C., Jr., & Harbeson, M. M. An engineering approach to the standardization of Performance Evaluation Tests for Environmental Research (PETER). Proceedings of the 11th Annual Conference of the Environmental Design Research Association, Charleston, S.C., March, 1980.

Kennedy, R. S., Bittner, A. C., Jr., & Jones, M. B. The utility of commercially available television computer games for assessing performance and other applications. Preprints of the 51st Annual Scientific Meeting of the Aerospace Medical Association, Anaheim, CA, May 1980, 163-164.

McCafferty, D. B., Bittner, A. C., Jr., & Carter, R. C. Performance Evaluation Tests for Environmental Research (PETER): Auditory digit span task. Proceedings of the 24th Annual Meeting of the Human Factors Society, Los Angeles, October, 1980.

Paivio, A., Yuille, J. C., & Madigan, S. A. Concreteness, imagery, and meaningfulness values for 925 nouns. Journal of Experimental Psychology Monograph, 1968, 76, (1, Pt. 2).

Shannon, R. H. Task analytic approach to human performance battery development. Proceedings of the 24th Annual Meeting of the Human Factors Society, Los Angeles, October, 1980. (a)

Shannon, R. H. A factor analytic approach to determining stability of human performance. Proceedings of the 13th Annual Meeting of the Human Factors Association of Canada, Point Ideal, Ontario, Canada, September, 1980. (b)

Shepard, R. N., & Teghtsoonian, M. Retention of information under conditions approaching a steady state. Journal of Experimental Psychology, 1961, 62, 302-309.

Thomas, D. J., Majewski, P. L., Ewing, C. L., & Gilbert, M. S. Medical Qualification Procedures for Hazardous-Duty Aeromedical Research. (Conference Proceedings No. 231, A3, pp. 1-13, 1978) London: AGARD, 1977.

Thorndike, E. L. & Lorge, I. The teacher's word book of 30,000 words. New York: Teachers College, Bureau of Publications, 1944.

Underwood, B. J., Borach, R. F., & Malmi, R. A. The composition of episodic memory. (ONR Contract No. N00014-76-C-0270) Evanston, Illinois: Northwestern University, May 1977. (NTIS No. AD A040696).

PROCEEDINGS OF THE SEVENTH PSYCHOLOGY IN THE DOD SYMPOSIUM
USAF ACADEMY, COLORADO SPRINGS, CO 16-18 APRIL 1980

Performance Evaluation Tests for Environmental Research (PETER): Interference Susceptibility Test (IST)

Michele Krause and Robert S. Kennedy
Naval Aerospace Medical Research Laboratory Detachment
New Orleans, Louisiana

Abstract

A program designed to develop Performance Evaluation Tests for Environmental Research (PETER) is in progress. Underwood's (1977) Interference Susceptibility Test (IST) was evaluated for inclusion in PETER on the basis of its suitability for repeated administrations. Baseline testing consisted of alternate forms of the IST being administered to 23 subjects for 15 workdays. The results show the mean of the total percent correct score continues to exhibit a slow increase over the entire experiment, with the standard deviation remaining constant subsequent to Day 7. Reliability correlations appear differentially stable after some training ($\bar{r}$ .75). The slope score, the traditional measure of IST, is unreliable, although the standard deviations are relatively constant. The total percent correct score is recommended for possible inclusion in PETER.

The Navy is developing Performance Evaluation Tests for Environmental Research (PETER) at its medical laboratory in New Orleans. The goal of the PETER program is to develop a multiple administration test battery which will be effective in detecting performance decrements that are caused by ship motion. Additionally, due to its nature, the test battery is expected to lend itself to the study of other stressors, such as toxic drugs, extreme temperatures and high pressure. The current phase of this project involves repeated testings of cognitive, perceptual and psychomotor tasks. In choosing a task for study, one or more of the following criteria must have been met: (a) performance has been shown to be disrupted in a thermal, inertial or hyperbaric environment, (b) it has been acknowledged to assess cognitive, information-processing, or memory functions, or (c) normal subjects have been distinguished from brain damaged persons (Kennedy & Bittner, 1977). One of the tasks selected for study was Underwood's Interference Susceptibility Test (IST) (Underwood, Boruch & Malmi, 1977). This task was originally designed by Underwood to study the effects of proactive interference. In this original study, 200 college students were tested on 24 separate tasks. Fernandes and Rose (1978) included the test in their studies of an information-processing approach to performance assessment. It is suspected that the more basic memory tasks which have been studied at NAMRLD (e.g. recall and recognition tasks) do not distinguish memory capacities in the same way as IST does. The Interference Susceptibility Test required associations to be formed, dismissed, and then new, conflicting associations formed during

The opinions are those of the authors and do not necessarily reflect those of the Department of the Navy.
This research was performed under Navy Work Unit No. MF58.524.002-5027.
The authors are indebted to Andrew Rose for providing stimulus material.

exposure to persons suffering from motion sickness, one of the authors found that "confusion" was reported as a frequent mental symptom. It is possible that IST is sensitive enough to measure a component of "confusion".

The purpose of the present study is to determine whether IST is suitable for use in environmental research. From our point of view, a task is considered suitable if it has task definition (i.e. differentiates between subjects) and is stable. In accordance with Jones (1979), stability exists when: (a) the daily group means asymptote or evidence a slight, constant slope, (b) day-to-day variance is constant, and (c) relative performance standings between subjects are constant from day to day. A recommendation of whether to include this test in subsequent PETER studies is based on these criteria. Reviews which describe this program in detail, as well as describe the results of previous tasks that have been administered, are available (Harbeson, Kennedy & Bittner, 1979; Kennedy & Bittner, 1977; Kennedy, Bittner, & Harbeson, 1980).

Method

Subjects

The subjects were a group of 23 volunteer enlisted Navy men, ages 19 to 24. To qualify for this medical research program, they had to be within the norms for Navy enlisted personnel in physical health, mental health and intelligence. All subjects were recruited, evaluated and employed in accordance with procedures specified in Secretary of the Navy Instruction 3900.39 and Bureau of Medicine and Surgery Instruction 3900.6. These instructions are based upon voluntary consent, and meet the provisions of prevailing national and international guidelines. A description of the subject selection procedure is given by Thomas, Majewski, Ewing and Gilbert (1978).

Task description

Stimulus material for each session was comprised of lists of trigram-digit pairs (e.g. NOB-2). A list was made up of five trigrams paired with digits from 1 to 5. During each session, three sets, each containing four lists, were administered. Across the four lists of each set, the same trigrams were paired with digits from 1 to 5, forming different combinations in each list. Stimulus material was provided by Rose. An example of stimulus material for one set is found in Table 1.

Apparatus and procedure.

Subjects were shown each of five trigram-digit pairs by means of a single slide, presented on a Kodak Ektagraph 450 AudioViewer^{R}. The rate of presentation was one slide every 3 seconds. A cueing slide appeared at the end of the list and at the beginning of the recall list. Each trigram was then shown by itself (in an order different from the paired presentation) for 4 seconds, and subjects recorded the number with which they thought each trigram had been paired. Subjects were tested in groups of four, at 8:00 in the morning, for 15 consecutive workdays.

Results

Two measures were taken across sets for four lists: (a) slope of lists and (b) percent correct for each list. In addition, mean percent correct was obtained for each of three sets (summed over lists) and an aggregate mean (over sets and lists) was obtained in order to compare results with Underwood, et al. (1977).

Figure 1 shows the mean percent correct responses across sets for the four lists. As expected, performance declines with each successive list that is presented. The impression of a learning curve over days is observable across each list. The greatest improvement is seen in List 1 (33%). The reason for the anomalous scores on Day 6 is obscure. Standard deviations, as seen in Figure 2, are level and unremarkable.

Percent correct performance for each of the three sets (summed over lists) showed that subjects exhibit a slight advantage for later sets (not shown), although the differences are negligible. Mean performance for the three sets, across lists progresses from 50.1 on Day 1 to 71.8 on Day 15. The average percent correct in both this study and the Fernandes & Rose (1978) study was 65%. Underwood, et al. (1977) obtained an 85 percent correct average when this test was interdigitated with 23 other memory tests.

When Underwood et al. (1977) correlated total correct responses for Sets 1, 3, 5 with those same scores from Sets 2, 4, 6, they obtained a value of $r = .81$. This correlation between successive sets (i.e. split half) in Underwood's study is compared to a correlation of $r = .74$ between successive days (i.e. test-retest) in the present research, wherein the number of observations are the same for both calculations. There is no evidence that the reliabilities of the present data are different from those of Underwood et al. (1977) ($z = .72$, $p > .40$).

Tables 2 and 3 show reliabilities within Lists 2 and 4. Because Lists 1 and 3 revealed comparable results, they are not shown. These correlations reveal that average percent correct performance appears to stabilize around Day 8. This result is, perhaps, more clearly illustrated when Table 2 is graphed as in Figure 3. This figure presents correlations of percent correct performance for selected testing days in a left-justified manner, enabling examination of all subsequent testing days. Although a progression towards stabilization occurs, the task definition remains too low to be satisfactory (Jones, 1979).

Figure 4 shows the means and standard deviations for the slope scores over lists. Mean slopes are variable and show no systematic trend. The standard deviations are equal to the means suggesting substantial differences between subjects. Table 4 shows slope reliabilities. Composite reliability for this score is essentially zero ($r = .04$).

Discussion

Percent correct scores for the individual lists provide evidence for stabilization within the second week of testing, but with task definition at too low a level to be considered useful. When the percent correct scores are summed over lists and sets task definition improves ($\bar{r} = .71$),

and reliabilities after Day 8 appear stable. This aggregate score is the one favored by Underwood, et al. (1977), who found it to be correlated with the slope measure. While less defensible as a measure of interference susceptibility, the percent correct score over lists and sets meets the minimum requirements for suitability for PETER and will be employed in subsequent analyses at this laboratory. It should be noted that the test in its present form, requires ten minutes to complete and yields a composite reliability in List 2 (as an example) of r = .53 Using the Spearman-Brown adjustment formula (Allen & Yen, 1979), reliability raises to r = .69 if the testing length is doubled. The total aggregate score improved from r = .71 to r = .83.

The chief finding in this experiment is that the slope score, theoretically the most meaningful measure of the interference factor, is unreliable (r = .04). This poor reliability over sessions is not due to insufficient variance between subjects and it occurred despite the fact that the slope means and standard deviations are stable. Fernandes and Rose (1978) also obtained low reliability for the slope measure (r = .05). It is probable that the same cautions which are associated with difference scores (Cronbach & Furby, 1970) may apply to slopes. Those authors suggest, as an alternative, analyzing the most complex condition with the simpliest condition as a covariate (in this case, List 4 with List 1 as a covariate). This analysis will be performed on the IST data at a later date.

In conclusion, IST as analyzed up to this point, is not an ideal candidate for inclusion in future PETER studies. It is recognized though, that with some modifications to the adminstration procedure, this test may reveal a unique factor of memory that would be useful to include in the final PETER battery. It may prove to be necessary, when studying other environmental stressors, (specifically impact acceleration) to place heavier emphasis on memory tasks because of the close connection between memory and other human systems and functions.

References

Allen, M. J. & Yen, W. M. Introduction to Measurement Theory. Belmont, California: Wadsworth, Inc., 1979.

Cronbach, L. J. & Furby, L. How should we measure "change" - or should we? Psychological Bulletin, 1970, 74, 68-80.

Fernandes, K. & Rose, A. M. An Information Processing Approach to Performance Assessment: II. An Investigation of Encoding and Retrieval Processes in Memory. (Tech. Report IAR 58500-11/78 TR). Washington, D.C.: American Institutes for Research, November, 1978.

Harbeson, M. M., Kennedy, R. S., & Bittner, Jr., A. C. A comparison of the Stroop Test to other tasks for studies of environmental stress. Proceedings of the 12th Annual Meeting of the Human Factors Association of Canada, Bracebridge, Ontario, Canada, 6-8 September, 1979.

Jones, M. B. Stabilization and Task Definition in a Performance Test Battery. (NAMRL Monograph No. 27). Pensacola, FL: U. S. Naval Aerospace Medical Research Laboratory, 1980.

Kennedy, R. S. & Bittner, A. C., Jr. The development of a Navy Performance Evaluation Test for Environmental Research (PETER). In, Productivity Enhancement: Personnel Performance Assessment in Navy Systems. Symposium presented at the Naval Personnel Research and Development Center, San Diego, CA, 12-14 October 1977. (NTIS No. AD 056047).

Kennedy, R. S. & Bittner, Jr., A. C. Progress in the analysis of Performance Evaluation Tests for Environmental Research (PETER). Proceedings of the 22nd Annual Meeting of the Human Factors Society, Detroit, Michigan, October, 1978. (NTIS No. AD A060676)

Kennedy, R. S., Bittner, Jr., A. C. & Harbeson, M. M. An engineering approach to the standardization of Performance Evaluation Tests for Environmental Research (PETER). Proceedings of the 11th Annual Conference of the Environmental Design Research Association, Charleston, S. C., March, 1980.

Thomas, D. J., Majewski, P. L., Ewing, C. L. & Gilbert, M. S. Medical Qualification Procedures for Hazardous-Duty Aeromedical Research. (Conference Proceedings No. 231, A3, pp. 1-13, 1978) London: AGARD, 1977.

Underwood, B. J., Borach, R. F. & Malmi, R. A. The composition of episodic memory. (ONR Contract No. N00014-76-C-0270) Evanston, Illinois: Northwestern University, May 1977. (NTIS No. AD A040696)

Tables

Table 1
Stimulus Presentation

	List shown to Se	Probe shown to Se	Correct response
LIST 1	DOG - 5	WIN	1
	WIN - 1	PEG	4
	PEG - 4	DOG	5
	NEW - 3	NOB	2
		NEW	3
LIST 2	WIN - 5	PEG	2
	PEG - 2	NOB	3
	DOG - 4	NEW	1
	NEW - 1	DOG	4
	NOB - 3	WIN	5
LIST 3	NEW - 2	NEW	2
	NOB - 5	NOB	5
	DOG - 1	DOG	1
	WIN - 4	WIN	4
	PEG - 3	PEG	3
LIST 4	DOG - 3	NEW	5
	NEW - 5	WIN	1
	NOB - 4	PEG	2
	PEG - 1	NOB	4
	WIN - 2	DOG	3

Table 2
Mean Percent Correct
Reliabilities for List 2

Days	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	.32	.06	.42	.01	.07	.43	-.04	.15	.27	.32	.25	.12	-.05	.24
2		.31	.46	.46	.33	.44	.30	.38	.41	.06	.24	.19	.22	.30
3			.62	.28	.00	.53	.58	.27	.68	.57	.50	.72	.70	.64
4				.35	.34	.70	.61	.49	.70	.45	.54	.61	.63	.56
5					.24	.17	.12	.46	.35	.23	.26	.45	.32	.30
6						.53	.24	.51	.41	.10	.28	.28	.05	.24
7							.46	.52	.68	.45	.67	.52	.51	.51
8								.55	.70	.53	.45	.41	.53	.47
9									.70	.44	.51	.51	.46	.32
10										.61	.68	.66	.58	.62
11											.64	.68	.67	.75
12												.53	.55	.46
13													.77	.70
14														.72
15														

Table 3
Mean Percent Correct
Reliabilities for List 4

Days	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	.32	.59	.40	.38	-.16	.34	.08	.60	.33	.32	.58	.39	.48	.44
2		.05	.10	-.07	.18	.43	.08	.15	.52	.42	.35	.16	.14	.20
3			.34	.18	-.43	.14	-.05	.16	.05	-.06	.34	.04	.06	.10
4				.33	-.01	.40	.23	.47	.25	.30	.15	.42	.51	.50
5					.23	.40	.48	.26	.48	.35	.24	.61	.55	.50
6						.05	-.00	.15	.06	.22	-.15	.16	.10	.27
7							.55	.35	.59	.81	.47	.42	.52	.46
8								.18	.34	.49	.09	.37	.55	.02
9									.40	.34	.45	.50	.64	.37
10										.44	.61	.60	.53	.49
11											.40	.55	.54	.46
12												.40	.41	.41
13													.73	.59
14														.44
15														

Table 4
Reliabilities of Mean
Slope of Lists Across Sets

Days	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	-.21	.15	-.14	.04	.03	.04	-.17	.16	.06	-.08	.08	.14	-.15	-.04
2		.10	-.02	-.13	-.05	.32	-.09	-.05	.10	.29	.09	.08	-.05	.02
3			.07	.16	-.10	-.05	.30	-.01	.32	-.31	.16	.01	.23	-.28
4				-.19	.20	.39	-.15	-.11	-.29	.01	-.06	.25	.19	.03
5					-.31	-.13	.34	.06	.42	.32	.40	.57	.56	.06
6						.32	-.12	.23	-.29	-.08	-.14	.07	-.09	.04
7							-.03	.28	-.08	.54	.40	.38	.03	.18
8								-.08	.24	.09	-.18	-.15	.26	-.15
9									.16	.08	.49	.29	-.01	.00
10										.01	.35	.17	.14	.33
11											.45	.42	-.11	.29
12												.60	.29	.14
13													.39	.07
14														.03

Figures

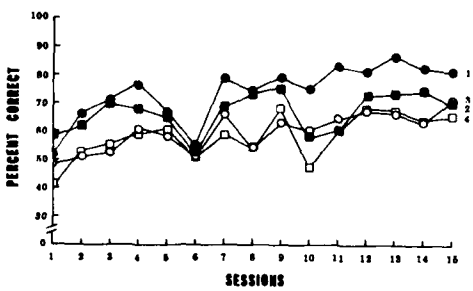


Figure 1. Mean Percent Correct Across Sets for Lists 1, 2, 3, & 4 Over 15 Days.

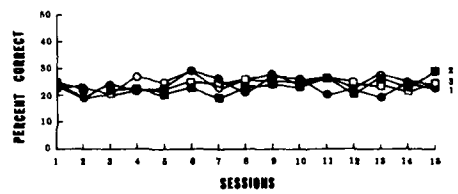


Figure 2. Standard Deviations of Percent Correct Across Sets for Lists 1, 2, 3 & 4 Over 15 Days.

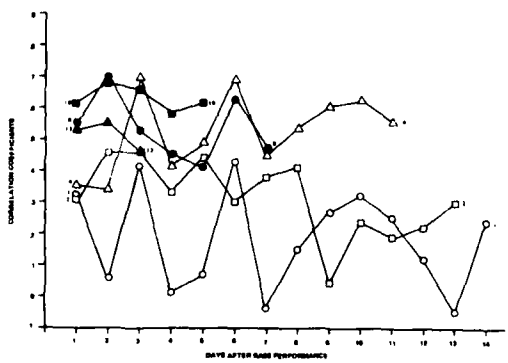


Figure 3. Reliabilities for Percent Correct Across Sets for List 2 for Selected Base Days 1, 2, 3, 4, 8, 10, 12 and Those Following over 15 Days.

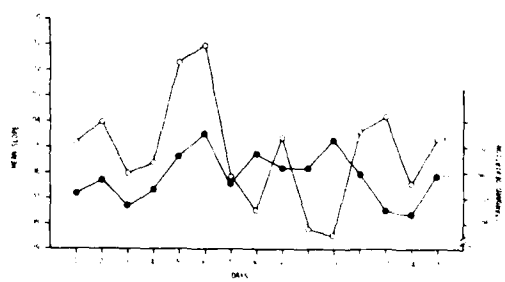


Figure 4. Means and Standard Deviations for the Mean Slope of Lists Across Sets Over 15 Days.

PROCEEDINGS OF THE 24TH ANNUAL MEETING OF THE HUMAN FACTORS SOCIETY
LOS ANGELES, CA, 13-17 OCTOBER 1980

ITEM RECOGNITION AS A PERFORMANCE EVALUATION TEST FOR ENVIRONMENTAL RESEARCH

Robert C. Carter, Robert S. Kennedy, Alvah C. Bittner, Jr., and Michele Krause
Naval Biodynamics Laboratory, New Orleans, LA 70189

ABSTRACT

Item Recognition (Sternberg, 1966) is a task which reflects the operation of human memory. This task was considered as a candidate for use in a battery of Performance Evaluation Tests for Environmental Research (PETER). Environmental research involves comparison of performances in a baseline environment and in a novel environment. It is desirable that scores be stable at different occasions in the baseline environment, so that changes due to the novel environment will be clear if they occur. It was found that item recognition results were similar to those obtained by other investigations, although the traditional item recognition score (slope) was unreliable across repeated measurements. The response time (RT) was stable for each of the four memory set sizes (1, 2, 3 & 4 items), from the standpoint of reliability, after the fourth session.

INTRODUCTION

Sternberg's (1966, 1975) item recognition task has recently been suggested for use as a performance evaluation test (Rose, 1974). If a test is to be used for environmental research, it must be administered repeatedly, usually to the same subjects, in a baseline condition and in the novel environment. It would be desirable for a test to provide unchanging scores in the baseline condition because any change associated with repeated measurement would be confounded with changes of performance due to the environment. Therefore, experiments are being conducted to determine whether tasks yield stable scores which qualify them for use as Performance Evaluation Tests for Environmental Research (Kennedy, Bittner & Harbeson, 1979). Jones (1980) suggests that stability is indicated when: (1) mean performance reaches nearly constant slope over time, (2) between subject variances are homogeneous over time, and (3) relative performance standings of the subjects, reflected in cross-session reliabilities, are constant over time. The latter two of these stability criteria, it is noteworthy, are sufficient requirements for simple repeated measures analysis of variance (Winer, 1971).

METHOD

Subjects were 21 Navy enlisted males meeting qualifications described by Thomas, Majewski, Ewing and Gilbert (1978). Testing was conducted once each day beginning on a Monday and continuing for fifteen consecutive weekdays. The test sessions lasted about 15 minutes per subject per day.

Subjects in this item recognition task were presented with a series of one to four digits called the positive set which were presented for 1 sec. per item. All other digits constituted the negative set. A probe digit followed presentation of the positive set by 2 sec. The subject was to select one of two responses depending upon whether the probe was from the positive or negative set. The duration from onset of the probe to the response was recorded as the response time (RT). Each session included ten trials for each positive set size of 1, 2, 3 or 4 unique digits. Half of these trials included probes from the positive set, and half were from the negative set. Within these restrictions the digits of the positive set and the probe digits were chosen at random, and were different on each day, but were the same for all subjects on any particular day. Daily means and standard deviations, and interday correlation (reliability) matrices (all calculated across subjects) were developed for each of the following scores: Mean RT's for positive set sizes 1, 2, 3, and 4; slope of mean RT versus set size; intercept of mean RT versus set size; and percent error. The slope and intercept scores for each subject on each day were computed by least squares regression. There was a regression equation for each subject which expressed the 40 RTs for that subject on that day as a linear function of positive set size. Slopes and intercepts from these equations represented individual differences, the reliabilities of which were shown in intertrial correlation matrices. Aggregate performance of all subjects on each day was summarized by averaging the subjects' slope or intercept scores.

Slope and intercept scores were calculated based on Sternberg's (1966) finding that RT increased linearly with positive set size. This finding has since been confirmed many times (Sternberg, 1975). The slope may be interpreted as the rate of search through short-term memory and the intercept is interpreted as time required for stimulus processing and response formulation (cf. Sternberg, 1966, 1975). These scores have been found to reflect differences among individuals' information processing capabilities associated with age (Anders, Fozard, & Lillyquist, 1972) and with aphasia (Swinney & Taylor, 1971).

RESULTS AND DISCUSSION

The present experiment differs from Sternberg's (1966) in that he reports results for "practiced" subjects while we show how the results are affected by the degree of experience. Our intercept score (450 msec) did not change appreciably during the experiment ($F(14,280) = 1.53$, $p > .1$) and is comparable to that reported by Sternberg (397.2 msec). However, our slope scores (Figure 1) decreased with practice ($F(14,280) = 5.32$, $p < .005$). This is a common finding (Kristofferson, 1972; Ross, 1970; and Simpson, 1972). Figure 1 indicates that the

slopes do not change very much after the third day of testing. Our average slope score on the third day (41.2 msec/item) is very similar to the average slope obtained by Sternberg with practiced subjects (37.9 msec/item). Our results contrast with Sternberg's in that our subjects' error rate was much greater than his (6% versus 1.3%); the error rate did not change with practice ($F(14,280) = .8$; $p > .3$).

Our main interest was to evaluate the use of the slope and intercept scores as measures of individual differences. Sternberg (1969) reported individual differences of slopes, which he conjectured to be related to different strategies of memory scanning. We too obtained significant individual differences of slopes ($F(20,280) = 2.57$, $p < .005$) and intercepts ($F(20,280) = 14.25$, $p < .005$). The cross-session reliabilities of these slope and intercept scores indicate the degree to which the scores represent enduring abilities. Figure 2 illustrates selected cross-session reliabilities of the slope scores. This figure shows the extent to which subjects' scores tended to remain in the same relationship to each other from day-to-day. The complete set of cross-session reliabilities for slopes are shown in Table 1. The reliabilities are uniformly low, and if they do stabilize, it is at a uselessly low level. Similar results were obtained for the intercept scores. The poor reliabilities cast doubt upon the potential of these scores for measurement of individual differences and they would make the test relatively insensitive to environmental effects.

In contrast, the reliabilities of the RTs from which the slopes are calculated are relatively high, being generally greater than $r = .70$. Figure 3 shows cross-session reliabilities of RT for positive set size 4 (RT4). (Similar results were obtained for other positive set sizes). These reliabilities stabilize after Day 3 and are substantial enough to differentiate individuals ($\bar{r} = .80$). The complete set of cross-session reliabilities for the 4-item RTs are shown in Table 2. Unfortunately, the RTs are not as meaningful as the slopes and intercepts. For instance, the slope is supposed to represent the rate of memory scanning. But does it? Figure 4 shows the mean reaction time to positive set sizes 1 through 4 on each day of the experiment. If the rate model were appropriate, then RT2-RT1 = RT3-RT2 = RT4-RT3. Clearly this is not the case. The interval between RT1 and RT2 is usually greater than any of the others. Perhaps the slope is unreliable because the rate it is supposed to represent is a fiction. Numerous authors have found, as we did, that the RT versus positive set size curve is nonlinear (Simpson, 1972; Kristofferson, 1972; Swanson, 1974; Juola & Atkinson, 1971; Ross, 1970). In our case, the nonlinearity cannot be explained as due to a time-error tradeoff because error rate was independent of positive set size ($F(3,60) = .16$, $p > .5$). Fitting a line to such data adds a bias (Draper & Smith, 1966) component to the error of the fit. Reliability is the ratio of the true variance to the sum of error plus true variance. Inflation of the error by the bias would cause the reliability ratio to collapse, as it did for the slope scores in this experiment.

Even though the increase of RT with memory load is not linear, it is still meaningful to think of the increment of RT when positive set size is increased. For instance, if RT4-RT1 were different from one person to another or if it were altered by a change in the environment, then we could infer a difference in the amount of time required to mentally compare the second, third and fourth members of the positive set with the probe. The estimate of RT4-RT1 could be improved by accounting for the covariance of RT4 and RT1 (Cronbach & Furby, 1965). This refined estimate of the time required for mental scanning could come from an Analysis of Covariance of RT4, with RT1 as the covariate.

SUMMARY

Sternberg's (1966) item recognition task has been scrutinized as a candidate performance evaluation test for environmental research. Sternberg and others (cf., Sternberg, 1975) have interpreted the slope of RT versus positive set size to reflect the rate of memory scanning during recognition. Our results are similar to those of others who have studied this memory scanning slope, except that we have calculated cross-session reliabilities for repeated measurements of subjects' memory scanning speed. The reliabilities are vanishingly small, indicating either that a person's memory scanning rate is changeable (and hence, of little use as an individual difference parameter), or that the slope score is a poor way to represent memory scanning rate. The later interpretation is supported by the finding that RT (especially for large positive set size) is an extremely stable score which also reflects memory scanning rate. RT for a large positive set size, with RT1 as a covariate, is recommended for further consideration as a performance evaluation test which represents memory scanning speed during environmental research.

REFERENCES

Anders, T. R., Fozard, J. L., and Lillyquist, T. D. Effects of age upon retrieved from short-term memory. Developmental Psychology, 1972, 6, 214-217.

Cronbach, L. J., and Furby, L. How we should measure change-or should we? Psychological Bulletin, 1970, 74, 68-80.

Draper, N. R., and Smith, H. Applied regression analysis. New York: John Wiley, 1966.

Jones, B. Stabilization and task definition in a performance test battery. (NBDL Monograph No. M-0001) New Orleans, LA: Naval Biodynamics Laboratory, 1980.

Juola, J. F., and Atkinson, R. C. Memory scanning for words versus categories. Journal of Verbal Learning and Verbal Behavior, 1971, 10, 522-527.

Kennedy, R. S., Bittner, Jr., A. C., and Harbeson, M. M. An Engineering approach to the standardization of Performance Evaluation Tests for Environmental Research (PETER). Proceedings of the 11th Annual Conference of the Environmental Design and Research Association (EDRA). Charleston, SC, March, 1980.

Kristofferson, M. W. When item recognition and visual search functions are similar. Perception and Psychophysics, 1972, 12, 379-384.

Rose, A. M. Human information processing: An assessment and research battery. Ann Arbor: The University of Michigan, 1974. (Technical Report No. 46)

Ross, J. Extended practice with a single character classification task. Perception and Psychophysics, 1970, 8, 276-278.

Sternberg, S. High speed scanning in human memory. Science, 1966, 153, 652-654.

Sternberg, S. Memory scanning: Mental processes revealed by reaction-time experiments. American Scientist, 1969, 57, 421-457.

Sternberg, S. Memory scanning: New findings and current controversies. Quarterly Journal of Experimental Psychology, 1975, 27, 1-32.

Swanson, J. M. The neglected negative set. Journal of Experimental Psychology, 1974, 103, 1019-1026.

Swinney, D. A., and Taylor, O. L. Short-term memory recognition search in aphasics. Journal of Speech and Hearing Research, 1971, 14, 578-588.

Thomas, D. J., Majewski, P. L., Ewing, C. L., and Gilbert, N. S. Medical qualification procedures for hazardous-duty aeromedical research. London: AGARD, 1977 (Conference Proceedings No. 231 A3 P. 1-13, 1978).

Winer, B. J. Statistical principles in experimental design (2nd ed.). New York: McGraw-Hill, 1971.

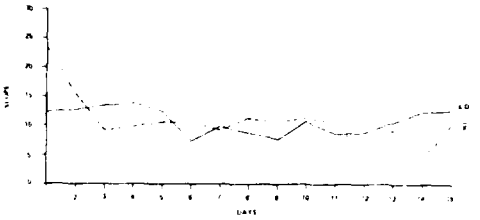


Figure 1. Item Recognition Slope Means ($\bar{X}$) and Standard Deviations (S.D.) Over 15 Days (N=21).

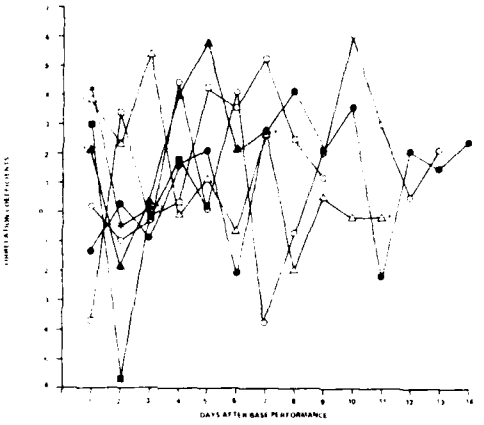


Figure 2. Item Recognition Slope Score: Intertrial Correlations Between Selected Days (1, 2, 4, 6, 8, 10, & 12) and Following Days (N=21).

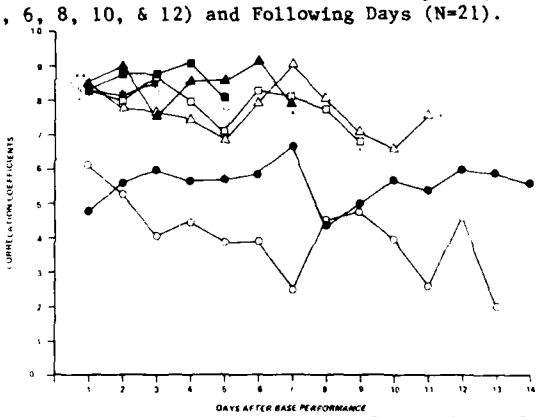


Figure 3. Item Recognition Time for Positive Set Size Four: Intertrial Correlations Between Selected Days (1, 2, 4, 6, 8, 10, & 12) and Following Days (N=21).

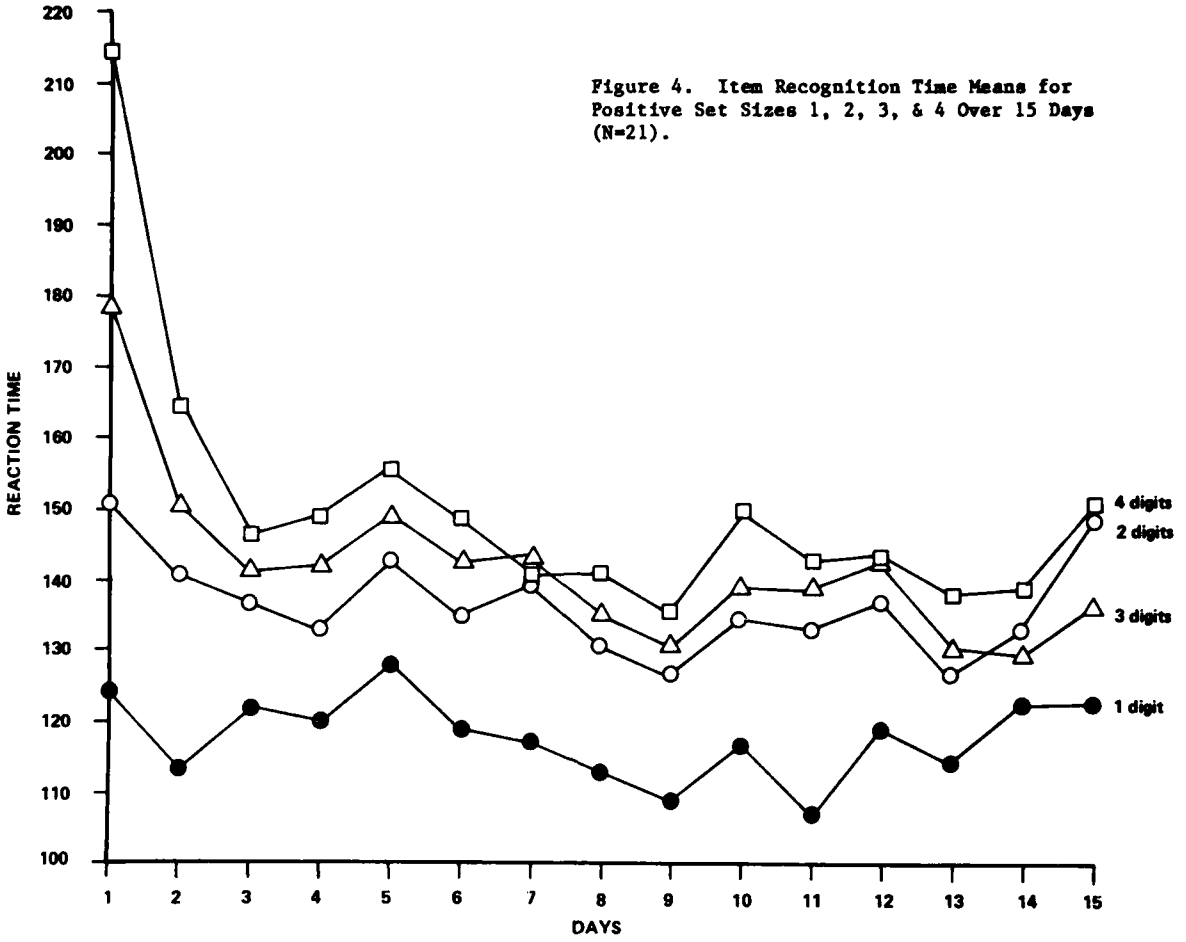


Figure 4. Item Recognition Time Means for Positive Set Sizes 1, 2, 3, & 4 Over 15 Days (N=21).

TABLE 1

Item Recognition: Slope Reliabilities over 15 Days (n=21)

DAYS	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	-13*	03	-09	16	21	-21	28	42	21	36	-22	21	15	24
2		02	-10	-03	44	01	41	-38	-07	19	61	31	05	21
3			-06	31	19	-17	-07	26	02	45	03	11	-43	11
4				39	23	54	01	12	-07	28	-20	05	-02	-02
5					31	02	-11	47	37	65	-32	37	04	-08
6						-37	34	-02	03	43	36	53	25	12
7							-07	-14	-05	04	-09	-24	-04	-07
8								21	-19	03	40	58	21	28
9									40	37	-56	25	09	03
10										30	-57	-02	19	01
11											-19	33	-03	28
12												42	-05	01
13													14	17
14														-09

* Decimal Points Omitted

TABLE 2

Item Recognition: Reliabilities for RT to positive set size 4 over 15 Days (n=21)

DAYS	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	48*	56	60	57	57	59	67	44	50	57	54	60	59	56
2		61	53	41	45	39	39	25	45	48	40	26	46	20
3			78	82	73	73	72	64	76	84	82	63	66	65
4				85	78	77	75	69	80	91	81	71	66	76
5					84	84	87	87	91	85	92	86	87	88
6						83	80	87	80	71	83	81	78	68
7							79	82	78	75	77	80	77	81
8								85	90	76	86	86	86	80
9									88	71	86	91	88	81
10										83	88	88	91	81
11											90	74	69	86
12												83	81	86
13													82	85
14														78

* Decimal Point Omitted


DAT
FILM